TURING 图灵新知

1930

1742

1200

825

20 影响数学发展的个大问题

[英] Tony Crilly 著
王耀杨 译



THE BIG QUESTIONS: MATHEMATICS



人民邮电出版社
POSTS & TELECOM PRESS

THE BIG QUESTIONS: MATHEMATICS

影响数学发展的20个大问题

“本书引领读者去探索影响数学发展的20个根本性问题，这些问题也正是各个时代的数学家们费尽心力去解决的。本书能让我们掌握当代数学的基本情况。”

——亚马逊读者评论

“20个大问题”系列的每本书均由相关领域的专家学者执笔，通过对某一学科里20个最根本的问题的解答和讨论，介绍了学科发展的历史，勾勒出了整个学科的概貌。针对每个精心挑选的问题，书中不仅做出详尽而富有启发性的解释，还不失时机地穿插了妙趣横生的历史故事，令读者在不忍释卷的阅读体验中畅游科学的奇妙世界。

本书是该系列的数学卷，通过20篇精炼的小文章阐述了数学中最重要、最深刻的一系列“大问题”，向大家充分展示了这门令人望而生畏却又至关重要的学科的魅力。

最奇怪的数是什么？无穷大到底有多大？数学怎样带来财富？一只蝴蝶扇动一下翅膀如何在地球的另一端引发一场风暴？如何破解复杂的密码？从这些看似寻常的问题入手，作者引领你进入数学的各个分支领域。这些知识点涵盖古今，内容妙趣横生，语言深入浅出，完美地结合了理论与故事，是帮助读者理解身边的现实世界的金钥匙。

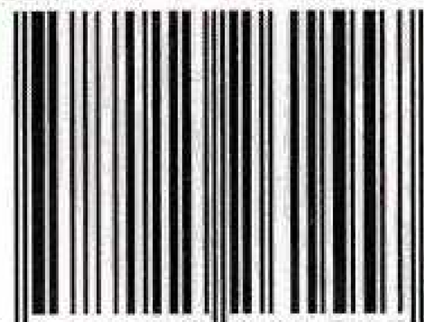




图灵社区：www.ituring.com.cn
反馈/投稿/推荐信箱：contact@turingbook.com
热线：(010)51095186转604

分类建议 科普读物/数学

人民邮电出版社网址：www.ptpress.com.cn



ISBN 978-7-115-26155-7

定价：29.00元

TURING 图灵新知

THE BIG QUESTIONS: MATHEMATICS

影响数学发展的20个大问题

[英] Tony Crilly 著

王耀杨 译



人民邮电出版社

北京

图书在版编目（CIP）数据

影响数学发展的20个大问题 /（英）克里利(Crilly,T.)著；王耀杨译. --北京：人民邮电出版社，2012.4
（图灵新知）
书名原文：The Big Questions:Mathematics
ISBN 978-7-115-26155-7

Ⅰ. ①影… Ⅱ. ①克… ②王… Ⅲ. ①数学—普及读物 Ⅳ. ①01-49

中国版本图书馆CIP数据核字（2011）第160155号

内容提要

这是一本数学科普书。作者通过20篇短文介绍了数学中最重要和深刻的一组“大问题”，同时也介绍了前辈学者的努力和成果，并指出了仍有待于深入探索的一些困难而诱人的未解之谜，内容涵盖数学发展史的方方面面，生动有趣，让读者为其深深吸引。

本书适合于对数学感兴趣的各个层次的读者阅读。

图灵新知

影响数学发展的20个大问题

◆ 著　[英] Tony Crilly
译　王耀杨
责任编辑　卢秀丽

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街14号
邮编　100061　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：880×1230　1/32
印张：6.875
字数：173千字　2012年4月第1版
印数：1－5 000册　2012年4月河北第1次印刷

著作权合同登记号　图字：01-2011-2425号

ISBN 978-7-115-26155-7

定价：29.00元

读者服务热线：(010)51095186转604　印装质量热线：(010)67129223

反盗版热线：(010)67171154

前　言

数学是一门我们都该有所了解的学问。学校的教学内容是一回事（而且不能打动所有人），但是这门学科的真正意义却是另一回事，她能提供的远不止于此。她不仅是辅助科学应用的一位默默奉献的助手，更与艺术有着根本性的联系。作为人类文明遗产的一部分，数学是鲜活的，并且正在持续地拓展着自身的疆界，而维持其生命力的，就是“大问题”。

数学中的大问题可谓多姿多彩。有些缘起于现代技术的巨大变革，有些则可以追溯至古代，时至今日仍余音不绝。有些已获得明确的答案，但又为一批新问题所取代，有些则依然故我，不为所动，历经若干世纪仍坚守前阵。答案的探求已处于哲学层面，也许注定无法以某种绝对的方式得到解决，但是问题本身依旧保持着迷人的风采。

这就是数学的特点。一个令人感兴趣的事实是：数学的发展是缓慢的。尽管在学校里，应对心算和各种技巧性小问题时速度很重要，但是在真正的数学研究工作中，绝对不存在单纯的快速所带来的优势。数学确实在前行，这是无可厚非的，但是它的步伐更像是熔岩流，柔缓并且不可阻挡，而少有伟大天才的“发现”时刻[①]。

数学具有一种与众不同的特质，使自己与其他科学相区分。一般的科学理论，一旦丧失了可信性，就会被抛弃。例如曾盛行一时的用来解释物质何以能燃烧的“燃素”学说，或者是用来解释光如何传播的“以太”学说。这

① 原文为“Eureka”，其含义接近于“我发现了”，传说阿基米德在洗澡时发现了浮力定律，喜极忘形，赤裸着跑上大街大呼“Eureka”。——译者注

些科学理论已被推翻，仅仅是出于人们好古的趣味而被列入科学史书籍之中。在数学中情况就不同了。一个已经证明正确的结果不可能在其后被证明是错误的，因而一个定理（即一个已经得到证明的数学事实）具有无限的生命。例如关于直角三角形的毕达哥拉斯定理[①]永远都是正确的。

今天的数学家可能不会再撰写研究论文，提出类似欧几里得在大约公元前300年时给出的那样的定理了。但是那些古代著述仍具有启示性，从基础文献中还可以发现新的思维方式。我们可以阅读古希腊数学家丢番图关于方程理论的著作，并且仍然可以从中获取教益，因为古希腊人所研究的某些类型的方程直至今日仍未获得解答。

这并不是说，时间对于数学理论和定理没有产生任何影响。我们经常对其加以修饰、提炼和裁剪，以适应当代的具体情况。数学的发展趋势是，当某些结论淹没在归纳总结的洪流中时，它们最终的命运不是进入垃圾箱，而是成为更具一般性的理论的注脚。

从数学的角度而言，我们所处的是一个激动人心的时代。新的问题必须要考虑到计算机时代的特征。这不仅是因为计算机能够高效地累加大量数字，更因为它们给我们对于数学证明的观念带来了冲击，进而引发关于数学本质的新问题。它们尤其擅长处理代数问题，以及显示几何形状和平面。

这里所考察的“大问题”，主要是指宏大的主题，并且集中探讨那些需要去关注的基本问题。这些问题会揭示数学从何而来，她有着怎样的发展历程，又将去往何方。它们都会给出答案，但这些答案绝不是已成定局的最终答案。它们引发了使数学家振奋激昂的新问题，并且告诉我们，数学是如何帮助我们理解自身居处其中的现实物理世界的。最重要的是，它们充分表明，数学是一门富于生命力的、活生生的学问。

---

① 我国习惯称为“勾股定理”，本着尊重原文的态度，此后仍直译为“毕达哥拉斯定理”。——译者注

数学披上今天在我们看来不那么可亲的外衣……当目睹这一切之后，你是否还会觉得数学是那么不近人情呢？

其次，“大问题”这个视角有可能帮助我们在更广阔的背景上理解数学知识。当我们在学校中学习数学知识时，会按照各种分支的名目来整理不同概念、定理，这个是几何的，那个是代数的，还有组合、分析等。看起来，数学知识就像分放在不同抽屉中的卡片，彼此分隔在孤立而漆黑的世界中。但是这可完全不是数学真正的样子。事实上，数学是统一的，她就像一张致密而牢固的网，不同节点之间都有着强韧有力的连接，只要将其中某些点提起，整张网就会随之而起，落入你的掌握之中。数学中的“大问题”就是这样的点。这些点可以自然地延伸开去，不断牵连出强有力的新方法和广阔的探索余地；换言之，它们也使数学获得了“繁衍能力”。

最后，我想谈谈这本书的读法。尽管数学阅读的方法绝不是三言两语能说清楚的，但是下面这个建议多半会是有帮助的，更要紧的是，它是容易实施的。这个方法很简单，就是“随时停下来”。比如，当介绍一个新问题及其含义的段落结束后，你可以先停下来，自己想想这个问题是否有什么可以着手尝试的方向；在看完一个问题的处理方法之后，文章中出现了“由此导致新的疑问”之类的句子，你也可以先不往下看，自己思考一下究竟还可以提出什么样的新问题。如果你的想法和下文内容是一致的，那就会收获一份满足感，继续看下去的动力更足了；要是你的想法和下文内容不一样，那就更有趣了，因为你可能想到了一些新的探索方向，因而从一个“接受者”转型为潜在的“创造者”。

作为译者，自己宣称所翻译的是一本好书大概不会是一件很有说服力的事情。如果我的这番介绍能够使你相信这本书的内容会有助于提高你对数学的兴趣，从而愿意去品味其中的妙趣，那么我将感到十分满足。

王耀杨

译者序

很多年前，当我还是学生的时候，经常听到数学老师说这样一句话："现在的孩子不会问问题！"时光荏苒，如今我自己也成为一名数学老师，我的同事们依旧不厌其烦地在重复着这句话。这着实是件很可疑的事情。

事实上，我相信自己逐渐明白了其中原委。"我国数学教育长期存在重教轻学，重标准答案而轻智力开发，重书本知识机械的接受而轻学生实践创新活动的倾向。同国外数学教育相比较，我国数学教育强调对数学课本上现成题目的解答而轻学生对数学问题的发现与提出问题的能力的培养。"（李兴贵等主编，《新课程数学阅读教学新论》，P129）

现在大家所看到的这本小书，就是一本和你谈"数学问题"的书。它要谈的事情包括：好的数学问题究竟应该是什么样的？怎样做到在探索过程中不断追问好的问题？如何看待探索问题时遇到的种种困难？等等。它的指向很明确，就是"向大师而不是他们的学生学习"——当我们仔细审视数学中那些真正的"大问题"，欣赏那些"大人物"在其中展示出来的风采时，才会逐渐理解数学中评价好坏的标准。

当然，这本书中所说的"数学问题"，并不是我们平常所说的"练习题"哦。这里所谓的"数学问题"，是指数学发展历程中一类特殊的问题，它们对于数学知识的建立与发展具有至关重要的意义。

具体来说，这类问题有以下三个特点。首先，它们的原始形态并不难懂，最多只要简短的解释，就能使大多数人理解其含义。例如关于质数的

一系列问题就是这样，甚至像混沌理论这种新生主题也是如此。其次，这些问题必须是足够困难的。或者说，它需要数学家作为一个整体付出大量的努力，有时可能会需要很多年的积累。数学家是一群喜欢困难的人，他们在面对难题时往往会兴奋多于紧张。因此，一个足够困难的问题常常会导致很多新方法的出现，有时也会出现“无心插柳柳成荫”的效果。最后但可能是最重要的是，这些问题还必须是“有繁殖力的”。也就是说，这个问题能够催生出大量的新成果，它就像一棵小苗，可以长成参天大树，还能结出累累硕果。曾经有人将这个特点描述为“非得要用十几个新问题才能解决一个老问题”，这个幽默的概括很好地揭示出事情的一个侧面；大数学家希尔伯特的比喻则似乎更切中肯綮地说明了什么是好的数学问题：一只“会下金蛋的老母鸡”。

那么，我们可以从欣赏这些“问题”中收获什么呢？或者退一步说，作者试图就这些“数学问题”和我们分享些什么呢？我想，至少有两个方面是值得谈一谈的。

首先，从“大问题”这个角度来看待数学，可以大大消除她原本可能给人带来的拒人千里之外的冷冰冰的印象。有这样一句经验之谈“每一个公式都会使一本科普书的读者减半”，这很可能与事实相差不远。仔细想想，至少从外观上看，数学是多么“不合群”啊！她使用极其简洁的符号语言，用非常严格的方式组织词句和段落，更喜欢用一种让人无可争议同时也多少有些无可奈何的方式去表达最终的结论……这一切的一切与我们常常看到的那些文学、艺术类作品相比，是多么不一样啊！但是，这真的就是数学的“本来面目”吗？至少不完全是。在这本书中，你会看到数学的另外一面：数学问题的最初来源常常是一些看似简单的现象，所使用的方法是我们几乎都会感到熟悉的，接着，在某个重要的岔路口，因为某些我们这些“局外人”也觉得可以接受的理由，她逐渐“转型”，变得越来越困难，而乐在其中的数学家们则运用类比、联想等看起来一点也不严格的方式去进行各种各样的尝试，直到他们得到某些足以打动自己的东西，最后才为

目　录

1 数学的意义是什么？——关于目标和前景的介绍

到了21世纪，数学已成为一门广阔而多层面的学科。她涵盖的活动类别是如此宽广，看起来几乎不可能将其种种表现归类于单一学科之中。一方面，她界定了诸如计数、时间和金钱等使日常生活得以运转的事务的基本要点，而另一方面，她看来就像一个密封的世界，在那里，一些不食人间烟火的伟大头脑炮制着无比复杂的谜题——接着他们再经年累月地尝试着去解决那些谜题。与此同时，我们的政治家们一如既往地宣称着：我们需要更多的数学家。那么，所有这些数学究竟有什么意义，她又是如何融入我们这个世界的呢？

我们今天所看到的数学根植于早期计数文化，其源头可追溯至大约公元前3000年。当然，数学一开始时只是用来处理实际问题。诸如市场商贸、税款支付、土地丈量、仰观星辰和历法设计等问题都要应用到数字、计算和某些基础的几何知识。但是过了大约一千年，埃及人开始研究他们所使用的数字系统的性质，而不大考虑是否具有明显的应用价值。他们还出于好奇心与智力上的愉悦感而创造数学谜题，就像我们今天享受报纸上的那些数独游戏一样。数学开始关注自身，数学家由此而产生。

大约公元前500年时，古希腊人在数学方面实现了巨大的跨越，一种真正具有数学思想的文化繁荣发展起来。他们的著作影响了其后的各个时代，直到今天仍为我们所研究。数学被视为最有用处的，因而成为正统教育中的固有组成部分。毕达哥拉斯、柏拉图、阿基米德和欧几里得只是那些推崇数学并影响后世千百年的希腊先贤中的一部分代表人物而已。

基督教时代的前几个世纪是倒退时期，那些热衷于数学的人会发现他们被驱逐到了文化世界的边缘。大约在公元400年时，希波的圣奥古斯丁提出“一个好的基督徒应该提防数学家和那些作出空洞预言的人”，谴责他们签订了“与魔鬼之间的契约，去蒙昧人们的心灵，将人们束缚于地狱的枷锁之中”。在那个年代里，与数学家这个词紧密联系在一起的是占星术士的邪恶行径，人们认为数学在潜在意义上是邪恶的异端主张，这种猜忌使数学在很长时间里毫无进展。

在16世纪，哲学家弗朗西斯·培根哀叹“纯粹数学之出色用途”仍未为人透彻地理解，不过有一件事标志着情况开始好转，伽利略获得了帕多瓦大学的数学教授职位。伽利略与罗马天主教之间的冲突（即教廷对他的某些发现的抵制）表明，当时的人们对于数学以及数学对物理学和天文学的影响作用的接受程度仍是很有限的。但是到了17世纪晚期，一场数学与科学的变革发生了，主角是伊萨克·牛顿和他的同时代者，他们永久性地改变了文化世界中的力量平衡。18世纪末和19世纪初的浪漫主义者可能会指责这种新的世界观，威廉·布莱克也许会嘲讽牛顿，但是作为科学的语言，数学已前途无忧。19世纪，数学系在各地大学中陆续设立，大批新颖而有挑战性的研究著述不断涌现。数学从此得到普遍认可。

实用性与纯粹性

关于数学有一种广为人知的争论，即究竟是实际需求孕育了数学创造，还是新的数学知识给实际应用创造了机会。从历史的角度来看，对实用性的考虑是数学发展的驱动力，但是当这门学问的内在生命开始萌发后，“纯粹的”数学思维就有可能独自为新的应用创造空间。好的数学基本上都具有应用潜能，但是你绝对不知道应用的时刻会在什么时候来到。敏锐的洞见也许会在下个星期出现，但也可能沉寂达50年甚至500年之久。

在数学发展的历史长河中，遍布着纯数学理论找到实际应用的例子。古希腊人精心建立了圆锥曲线的理论，后来人们发现，这正是17世纪时约翰尼斯・开普勒与伊萨克・牛顿断言行星以椭圆轨道运行时所需要的工具。多维数组的理论（矩阵代数）是在19世纪50年代为处理数学内部问题而建立起来的，而它恰好就是70年后快速发展的量子理论中的“矩阵力学”所需要的。而当乔治・布尔建立一个将逻辑转化为代数（即布尔代数）的系统时，他也绝对无法想到，他为一个世纪之后的计算机编程提供了语言载体。

就在50年前，富于影响力的英国数学家哈代还曾写道，他在从事数学研究时不会受限于必须为其思想找到“实际用处”的想法。事实上，令他感到欣慰的是，那时的数论仍是远离实际应用的。但是今天，他可能已无法再称许这种隔离状态，因为在这个世界中，他的纯数学对于计算机安全领域来说具有极其重要的意义（见第14章和第20章）。今天我们有很多种关于维度的理论，但是当曼德尔布罗特在20世纪70年代中致力于“分形”研究时，大概很少有人会猜测出它们的潜在应用（见第12章）。

但是数学家确实也是在意需求的。在18世纪，詹姆斯・瓦特遇

到了如何将蒸汽机中活塞的直线运动转化为旋转运动的问题，其结果是工业革命期间诞生了几何联动理论。当第二次世界大战需要密码破解者时（见第14章），拥有非凡才能的数学家从众多大学应征而来，结果是世界上第一台电子计算机的建成。

因此，纯粹数学与应用数学之间始终保持着一种共生关系，在电子时代，这一点更是显得格外真实。没有数学，计算机将一无是处，数字摄影技术根本不会出现，手机也将进入沉寂的世界。但是今天，职业数学家的“纯粹”研究也将大大受益于计算机的计算能力：这次轮到“应用数学”反哺“纯粹数学”了。

数学还有自我意识的一面，即她在哲学层面上反思自身的一面。关于这方面的历史呈现出这样一种发展历程：从古希腊人所认为的数学家揭示的只是早已存在的真理，到关于数学家角色的一种更为精巧的定位，其中涉及创造性和想象力（见第19章）。

在现代数学中，知识发展的基础在于公理和逻辑演绎。古希腊人预设了他们的公理的真实性，而今天的数学家则只期望公理是相容的。20世纪30年代，库尔特·哥德尔给数学带来了冲击，他证明了“不完备性定理”，这个定理指出，在一个形式化的公理系统中，某些数学命题在只使用该系统公理时既不能证明也不能证伪。换言之，当今数学中可能存在着不可证明的真理，因而也许只能保持现状。

尽管现代数学可以说是广袤而丰富的，其根基仍可如学校课程中那样划分为算术、代数和几何等分支。那么，它们的核心是什么？它们又将向何处去呢？

数及其特性

在数学的保留节目单中，用以计数的数字始终保持着最为重要

的地位，它们是所有数学家的起点。它们的演变历史可谓丰富多彩（见第2章），毫无疑问，我们最终使用符号0~9来表示的“以10为基”的系统并不是必然的。比如，最开始时并没有0。

质数（只能被自身和1整除的数）的特性是非常让人着迷的研究对象。我们仍不清楚它们在计数数中是如何分布的，考虑到我们认识质数已超过2000年之久，这一点几乎是难以置信的（见第3章和第20章）。除了计数数和其中的质数之外，在几个世纪中这份节目单就扩展到还包括负数、分数以及通常所称的“无理数”，即无限不循环小数。所有这些合在一起，便是数学家所说的“实数”（见第4章）。

数的内容远不止于这些。“实数”还只是一维的。我们可以设想它们在数直线上向左方（即负数）和右方（即正数）伸展。当数学家们凭借我们所称的“复数”（见第5章）而勇闯二维世界时，一次伟大的飞跃来临了。它们在解方程和提供新的分析理论等方面为数学家带来了更强的力量。今天，“复数”对于诸如电和磁等现象的研究可谓至关重要。

于是，我们有了很多种类型的数，但是它们要延展到哪里才会是尽头？自古以来，数学家便注定要与无穷这一主题进行较量。自亚里士多德以来，人们就认为“潜无穷”是存在的——这是一种永远不可能到达的无穷。但是到了19世纪，格奥尔格·康托尔引入了另一种无穷的观念，使得我们有可能去讨论很多的无穷（见第6章）。

几何、代数与数学中的变革

两千多年以来，几何学一直受控于古希腊人那似乎是无可抗拒的权威之下，直至今天，学生在校园中所接受的多种规则也是由他们制定的。特别是欧几里得，他依靠自己缜密的思维能力建立了一套呈现为如经典般真理的几何学知识系统。但是随着时间流逝，欧

氏几何学逐渐暴露出不足，人们最终明白，还存在着其他有效的几何学可以用来处理二维、三维甚至更高维度的现象（见第7章），并产生了“流形”的概念——这种形状的局部和整体具有不同性质的几何学(见第17章)。这些几何学甚至比欧氏几何更有资格宣称是“宇宙的几何学”，那可是一个深受物理学家关注的主题。

物理学家利用几何学去追索事物和宇宙的奥秘，生物学家和医学研究者则采纳了另一种类型的几何学，“纽结理论”，以便理解DNA的拆解和分析——这项实践引发了关于DNA检测的争议，并且在人类身份识别与案件侦破等问题中产生了众多复杂而又难以预料的后果。总而言之，数学家为科学家提供了不同的几何学，在这个工具箱中，科学家们可以选择看起来最适合于处理当前工作的工具。

将几何学转译为代数的语言是一个重要的转折点，这项发展应归功于17世纪的笛卡儿。在20世纪，对称的几何学也化身为代数学。对称性是一种难以描述的性质，在数学中（以及在其他很多领域中）经常被用来定义美（见第15章），现在我们可以用数学的“群论”来把握它了。群的概念在近世代数中处于核心地位，它使得对称性可以在微观尺度上进行考察（见第18章）。在一项可以追溯至19世纪的宏大研究项目中，数学家们最终于1981年完成了有限单群的分类。在这个“巨型定理”中，人们创造出一幅关于群的地图，所有的群都被划分到各族之中，此外还有26个散在单群——其中最大的一个包含大约8×10^{53}个元素，也就是8后面接着53个0。现在群论在理论物理学中占据着重要的地位，因为空间的变换就构成了群；在化学和晶体学中也是如此，因为在这里对称性出场了。

在一个代数问题中“求出x的值”，这对于每个受过基本教育的数学界人士来说都是非常熟悉的。这类“逆”问题是数学所擅长的领域，其应用颇为广阔。在这里，我们经常需要求一个“未知量”，

但是最初我们所得到的只是一个关系或者是关于这个未知量的一个方程。比方说，已知将一个正方形的边长增加3米所得到的面积为400平方米，作为一个逆问题，我们就能够计算出最初那个区域的未知边长x。利用代数知识将方程$(x+3)^2=400$“展开”，即可得到$x=17$。当数学界的前辈们的研究成果为我们提供了一系列公式来完成这些工作时，我们也就乐得走捷径（见第11章）。

向太空发射火箭，就要用到“微分”方程，而这意味着“微积分”的使用（见第8章），这种方法的典型应用就是速率和加速度的测定。有各种特定类型的微分方程，都有已发展成熟的理论支撑，但是也存在着很多没有精确解的“一次性”[①]方程。庞加莱建立了微分方程理论的一个新分支，这是一种“定性”理论，它关注的是解的性质而不是如何求出具体的解。这一研究导致了“混沌”理论（见第13章），从而为新的拓扑理论提出了一个截然不同的方向，彻底背离了我们看待形状的方式（见第17章）。

新颖的和未知的数学

“拓扑学”对于只具有普通水平的非数学界人士来说可能不是那么容易理解，但是另外两个相对晚些的发展成果可能更为人所熟知：概率和统计。

数学中最出色的现代产物之一，概率理论（见第10章），使我们能够用定量方式来把握不确定性。17世纪时，娱乐性的数学以对赌博问题的分析为我们开启了这一理论，而现在，这个理论逐渐成熟并发展出严谨的微积分方法，从而成为风险分析的核心基础。统计

① 原文为“one-off”，指“一次性的”。这里的意思可能是说，无法给出公式解法的方程只能求出数值解，也就是每次针对具体数据只能求出一个解，而不是像公式解法那样，一个公式就给出了一大类方程的解。——译者注

学是与概率相关的一个领域（见第9章），它为如何恰当地处理数据提供了理论，也为进一步实验提供了背景。在某种意义上讲，统计学开端于农业实验，不过今天它已应用得极为广泛，以至于任一人类活动，从政治到医学，几乎没有不用到统计学的。

对统计学及其他数学成果的运用，很自然就会使人想进行预言活动，以便了解未来（见第16章）。人口统计学家想要对五年内的人口数目做出合理的预言，证券交易人则试图根据统计学证据和预示性信息来预测股票市场。如何做到这一点呢？这些都是很困难的问题，如同预测天气一样，要依赖于至今仍无法求解的数学方程（见第20章），而“蝴蝶效应”更使其难上加难（见第13章）。

因此，数学也有古老与新颖之分。为了避免认为已接近大功告成而只会坐享其成，我们应该随时提醒自己仍未解决的数学问题是存在的，而且数量很多（见第20章）。况且，若不这样做，数学便会逐渐枯萎。有一些尚未解决的大问题长年困扰着思考者们，例如哥德巴赫猜想和黎曼假设，这两个问题都是与质数有关的；不过还有一些值得注意的新问题。当然，目前已有相当进展，其中一些足以荣登头版。1994年，费马最后定理的解决使数学成为公众视野的焦点（见第15章）。在此之前，数学与计算机联手解决了“四色定理”问题（见第11章），以及最近，一位隐居世外的俄罗斯数学家震惊了全世界，他证明了百年问题庞加莱猜想——而且他甚至不愿去领取属于自己的一百万英镑奖金（第17章）。

那么，数学的意义究竟是什么？就某些方面而言这是个奇怪的问题。我们承认它们可以仅仅是活动，是思维过程和对想象力的训练，人们可以从中获得满足——人们总是沉溺于此，而且还将一直沉溺下去，因为他们必须这样做，要是有人想寻找它们的用处，就会发现它们的用处随处可见且不断激增。如果有人愿意去探索数学

对于我们关于世界、宇宙、自然以及人类交往的知识所贡献的一切，那么也会得到同样的感受。在改变生活的各个方面，数学可以做到和已经做到的事情是无可估量的。但归根结底，激励数学发展的是人类最基本的、足以作为其特征的品质——永不满足的好奇心。

2 数，从何而来？
——从骨头上的刻痕到十六进制

在日常生活中，我们和数密切相关。早上一醒来，我们使用朦胧的睡眼去关注钟面上的一圈数字，当然，更为常见的是数字闹钟上的闪烁符号；我们可能要乘坐134路公共汽车赶往大学，或者是追赶前往帕丁顿的08:32的列车去上班；我们要数好零钱去买午餐，我们要查看记事簿上的日期，我们要按动移动电话上的按钮；在一天快要结束时，我们会漫无目的地上下滚动那控制数字电视频道的令人眼花缭乱的箭头，直到我们最终上床睡觉时，还要最后再看一眼时间。数在我们的生活中是那样地根深蒂固，我们又是如此地沉浸于数的世界中，无法看清这个万能工具的庐山真面目。那么，它们究竟从何而来呢？

当然，我们所遇到的很多数只是单纯的名称或标签。理论上讲，用数以外的方式也能使我们搞清楚公交路线。著名的亨氏“57变”[①]和杰克·丹尼的“Old No.7”[②]巧妙地暗示着一系列具有微妙

① 1896年，亨氏提出了一个非常简短但极具吸引力的销售广告——“57变”，将当时它的60多种产品归为57类。这“57变”是指亨氏公司在一年52周内可以为顾客提供不同的食品，加上圣诞节、感恩节、新年、独立日和复活节5个节日的食品，顾客在一年中可以享用57种全新的佐餐食品。——编者注

② 杰克·丹尼（Jack Daniels）是世界十大名酒之一，杰克·丹尼酒厂1866年诞生于美国田纳西州，是美国第一间注册的蒸馏酒厂。该厂生产的威士忌酒瓶上标有“Old No.7”的字样，但对于它的具体含义却众说纷纭，其真相始终都是不解之谜。——编者注

差异的罐头食品和威士忌，但它们也还是人造标签——商标。即使是这些商标，它们的有效性也要依赖于人类社会为了将物体排序（第一，第二等）和计数而建立起来的数字系统。

今天，人类几乎已是普遍地采用由0到9这10个符号构成的同一数字系统来进行计数和排序。这些符号组合起来便能力非凡，足以表示从星系之间的辽阔距离到一个原子核的半径，而且可以用不同方式去表示它们。非专业人士可能会将地球到太阳的距离写成93 000 000或是九千三百万英里，而一个数学家或科学家则更欣赏9.3×10^7（即9.3乘以10的7次方）的简洁精巧。要描述一米的一千分之一，我们有3种方法：0.001米，1毫米和10^{-3}（即10的–3次方）米，但一个原子核的微小直径，则最好还是表示为10^{-15}米。

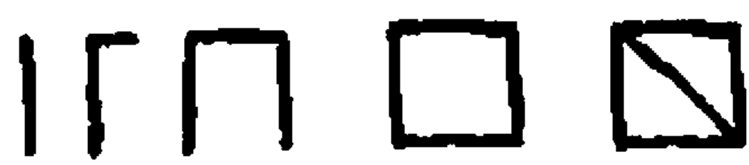
尽管那10个符号是如此精巧而威力无比，它们却绝对不是理所当然的，在人类发明数系的过程中，它们既不是开始，也不是结束。

最早的计数方法

研究人员已经发现了可以追溯至3万年前的物证来证明数字的早期记录形式——“计数棒”，即带有表明数量的标记的木棍。在非洲和东欧，用于记录数字的有凹痕的骨头进一步证明人们曾从事过计数活动。在13世纪的英格兰，征收税款时仍然要使用计数棒，更令人惊讶的是，这种传统方法一直持续到19世纪20年代才被纸质记录所取代。直到今天，当我们要记录持续增大的数量时，比如比赛中的得分或者是统计调查中收集的数据，计数标记仍然是有作用的。还有一种类似的方法，每计数到5就画一个像是篱笆墙的记号，它可能有着极为悠久的历史。



画篱笆墙的计数系统


一种南美的计数法

全世界都可以看到标记方法的使用。比如一种来自南美洲的方法，也是采用五条线，不过画出的是不同的图形。

在澳大利亚土著居民的原始计数系统中，数量超过小额数字后便混为一谈。在塔斯马尼亚，人们使用一种可以翻译为“一个”、“两个”和“多个”的计数系统。而在大陆地区的昆士兰州，曾使用一种包括“一个”、“两个”、“一个和两个”、“两倍的两个”和“多个”的计数系统。

巴比伦人和埃及人

一种真正意义上的数字系统在“文明的发源地”（中东地区）出现了。巴比伦文明在美索不达米亚地区繁荣发展起来，这一地区位于底格里斯河和幼发拉底河流域，是今天伊拉克的一部分，它的首都巴比伦坐落在今天巴格达城以南80千米处。在公元前3000年，古巴比伦人使用一种围绕数字60而建立的数字系统。我们今天的文化生活中还保持着它的残留痕迹，比如在时间的度量中一分钟有60秒，一小时有60分钟；另外由于巴比伦人将数学应用于他们的天文学中，因此一个完整圆周的度数也为360。基于数字60而建立的数字系统在数学方面是有优点的，比如，60能够被多达11个比自己小的数整除，即1, 2, 3, 4, 5, 6, 10, 12, 15, 20和30，对处理数量分配问题具有明显优势。


两个巴比伦数字符号


巴比伦数字符号中的数字23

数字1~60只需用两种符号即可表示出来，这两种符号很容易刻在用来记录的泥板上：它们是一条竖线和一个楔形符号。数字1用竖线符号表示，2~59则分别用两种符号的不同组合形式来表示。

巴比伦人从左到右阅读文字，他们的数字也是根据在一行中的相对位置来确定的，也就是说，和今天的数字体系是一样的。数到60后，他们便重新开始，用一道竖线符号来表示60，就像写1时那样，因此只有结合上下文才能解释其实际含义，比如要度量的对象是角，那就更可能是60度而不大可能是1度。他们没有相当于零的符号。

古埃及人建立了一种完全不同又非常复杂的数字系统，并且充分地呈现于建筑金字塔这一超凡的实践活动之中——这一壮举需要三维几何的知识和相当精确的度量水平。这一系统大约在公元前2700年左右出现，是以10为基数的。对于数字1~9，他们使用竖线计数符号，接着对数字10和100赋予两个不同的符号。与巴比伦人不同，埃及人是从右到左书写的。表示较大数的符号要更繁复一些，例如，他们用小鸟符号来表示数字100 000。



表示10和100的古埃及数字符号



古埃及符号系统中的数字234

埃及人的数学主要是与实际事务相联系的，但是他们也掌握某些精妙的算术技巧。他们做乘法的方法是很特别的。在我们现代知识系统中，心算乘法需要熟悉乘法表，我们往往是在学校中借助机械记忆的方式而掌握的。但是，生活在古埃及时代的孩子们实际上只需要乘以二的乘法表，因为他们做乘法的方法需要使用“算盘”[①]。

吠陀雅利安人

让我们继续东行，在公元前二千年时，吠陀雅利安文明从中亚

① 根据古希腊历史学家希罗多德的报道，古埃及人是在圆盘上用小石子进行演算操作的。——译者注

扩展到印度的印度河谷，关于它所使用的算术方法的记录在大约公元前1000年就出现了。在名为《吠陀》的古印度文献中包含着诗歌、文学和古训等内容，其中可以看到这种文明的19条数学“箴言”，或者说是文字公式。这些箴言为处理一系列算术问题提供了巧妙解法，有的还有多种解法。例如，其中一条“竖直与交叉”箴言是关于整数乘法的——在现代数字系统中这需要一段冗长的心算过程。它是通过一系列乘法和加法过程来实现的。比如我们想要用13乘以24（假设我们使用熟悉的符号），首先画一个网格，上面写13，下面写24。

1	3
2	4

吠陀雅利安式网格

将“竖直”数字相乘，再排列在一起，得到数212：即，(1×2) & $(3\times4)\to2$ & 12，记作212。接着将对角线上的数“交叉”相乘并加在一起，所得结果乘以10：即，$(1\times4)+(2\times3)\to4+6=10$，乘以10得100。

现在，将212和100加起来，就得到正确结果312。这种方法看起来像是魔术一样，而它的正确性就在于如下原理：要将两个数相乘，比如说$ab\times cd$，实际上就是进行乘法$(10a+b)\times(10c+d)$。

零的出现

零是数学符号中的后来者。它有两种主要功能，一是作为占位符号而出现，例如可以用来区分数字27和207。无论是巴比伦人、埃及人、希腊人还是罗马人都没有相当于零的符号——他们没有表示“没有”的符号。零的第二个功能是作为一个真正的数，在这个意义

上它的起源可以追溯到印度数学家婆罗摩笈多，他在大约公元600年时尝试将零引入了数字系统中。

零的拉丁名称（cifra）逐渐演变为zefro，在意大利文中就变为zero。法语中的“chiffre”（即“zero”）译为英文时变成“cipher”，这个词今天已不常用。在英语中我们用“nought”（即“无”）来表示零的意思，它来自于“nothing”这个词。从数学角度来讲，“无”其实是一个错误的称呼，因为以为数0就是什么都没有本身就是错误的。零无疑是有意义的，这个符号在我们今天的观念中发挥出真正的巨大力量。正如婆罗摩笈多所认识到的那样，在数学意义上的困难是，如何协调新来到的数字0与数系中其他成员之间的关系。

基数与二十进制的残留痕迹

数学家们会谈论在不同文化的计数系统中具有基础性地位的“基”。我们可以将它们理解为计数系统的核心单位或者是搭建系统所用的砖石。今天我们所用数系的基是10，即十进制系统，呈现为0~9这10个符号。但是历史已经见证，不同文化为了让数字适应其不同需要而分别采用过2, 3, 4, 5, 12, 20作为基，当然，还要算上巴比伦人使用过的60。即使到了今天，在我们的数系中依然可以看到继承不同文明之成果的痕迹，证据就是我们描述数时所用的语言。例如，在德语和英语中，“eleven”（德语为elf）和“twelve”（德语为zwölf）这两个词，从语言学的角度来看显得很奇怪，它们在10以前的各个数字和紧随其后由“ten”衍生出来的表示“十几”的数词之间形成了一个断层[①]。它们是以12为基数之数系的残留痕迹，实际上就和在使用十进制之前以12便士为1先令是一个道理。

① 从形式上看，从“十三”（thirteen）到“十九”（nineteen）的英文词的构词法是一致的，都是将个位数字与十位数字连接起来，只有一点细微变化。相比之下，“十一”和“十二”这两个词的形式显得与众不同。——译者注

在今天的危地马拉和墨西哥东南部地区，在被征服之前的玛雅文明中，曾采用过一种以20为基的数系，我们称其为“二十进制”系统。(有趣的是，玛雅人也是最早拥有以位值制为基础且包含零的概念的数字符号系统的民族之一。) 这种数系的来源常常被归因于我们的10个手指和10个脚趾，很多文化中都包含着以20为基的数系的残留痕迹。在英语中（以及在法语和德语的相应词汇中），表示“十几”的数词序列总会在十九停下来，以便为特殊的数词“二十”让路。

> “造物主的手法神秘莫测。但是他使用以十为基的数字系统，而且喜欢凑整数。”
>
> 斯科特·亚当斯，美国漫画家

20这个量还以其他很多种方式出现在我们的语言和生活之中。它有一个同义词是“score”，说到这个词，大家可能马上会想到英国国教祈祷书中将寿命70年称为“3个score又10年”。英制度量衡中包含英担（cwt）这个单位，20个英担就是一吨；在使用十进制英镑货币单位（1971年）之前，1英镑等于20先令。法语中也残留着曾使用二十进制系统的痕迹，有一个不是很常用的词，表示八十——quatre vingt（四个二十）。

十进制

尽管有各种关于其他数系的残留痕迹，今天这个世界统一使用的还是以十为基的“十进制”系统。

从很多角度来讲，这对于人类是很自然的选择：我们有10个可以用来计数的手指。古罗马人使用以10为基的系统进行整数计数，不过对于比较简单的计算，他们使用以12为基的分数系统，因为12可以被2, 3, 4和6整除。(某些历史学家认为，之所以用12作为基，是因为我们的每根手指上有3个关节，要是不包括拇指的话，每只手上

就有12个关节。）事实上，罗马人对于数学发展的贡献可谓微乎其微。但拉丁语是很多种现代欧洲语言的根源，因此在很大程度上讲，我们用来表示数的词是从拉丁文来的；而罗马人的数字书写系统，即I, II, III, IV, V, …, X等，则与我们所使用的“正常”数字平行地持续发展，特别是在记录日期的时候。

今天我们使用的数字符号最早是由婆罗摩笈多时代的印度数学家所使用，并由阿拉伯学者继承而来。随着阿拉伯旅者、商人和征服者的足迹拓展到北非并进入伊比利亚半岛，这些符号也随之传播开来。到12世纪时，阿拉伯人的数学知识也传播到西方来：9世纪数学家花拉子密的《论用印度数字进行计算》以拉丁文译本的形式出现，比萨的莱昂纳多（也称斐波那契）在他1202年出版的《计算之书》中大力宣扬这种印度-阿拉伯计数系统。13世纪时，英国哲学家、数学家和异端修道士罗吉尔·培根曾记录过这种符号。

罗吉尔·培根记录的数字符号（13世纪）

经过很少的一点改动后，到16世纪时，普遍使用的10个数字符号就已和我们今天所用的非常相似，并且随着大批量印刷技术的出现而获得了更高的标准化程度。

二进制

尽管事实已经证明十进制系统是强有力的也是善于适应变化的，但现代计算机技术却还是形成了一种不同类型的数字系统。由于计算机中的每个开关处于或“开”或“关”的状态，计算机技术就建立在识别这两种状态的基础上。由此产生出二进制系统，它的

字母表中只有0和1两种符号。由于我们的头脑并不会本能地用二进制数进行思考，因此我们需要一种将十进制数与二进制数进行相互转化的方法。

十进制数	二进制数
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001

十进制—二进制转换表

在十进制系统中，我们使用10的幂次来表示数，例如数312是由3个100、1个10和2个1组成的。而对于二进制系统，我们必须用2的幂次来表示数，也就是每次翻一倍：1, 2, 4, 8, 16, 32, 64, 128, 256, 512等。

因此，用划分的方法可以将数312转化为二进制形式，首先找到不超过312的2的最高次幂，即256，接着逐次向下。由此我们得到312 = 256 + 32 + 16 + 8。但是接下来，要得到二进制的数，我们还必须考虑到那些在上述表达式中没有出现的2的幂次，在出现2的幂次的位置放一个“1”，没出现的位置放一个“0”。补充完整后，得到：

$$312 = 1 \times 256 + 0 \times 128 + 0 \times 64 + 1 \times 32 + 1 \times 16 + 1 \times 8 + 0 \times 4 + 0 \times 2 + 0 \times 1$$

换言之，十进制数312变为二进制数100 111 000，或者用脚标数字来指明所选择的基数：

$$312_{10} = 100\ 111\ 000_2$$

不过，还有一种由两步构成的技巧可以将十进制数转化为二进制数，它类似于古埃及人在做乘法时所使用的加倍方法。首先，将该数字置于表格的最右端，接着将它不断地除以2，直到结果是1为止，在整个过程中不用考虑任何余数。还以312为例进行说明，到某一步骤时我们要用39除以2；记录答案19，而忽略余数1。由此得到表格中

上面一行的内容。其次，在表格的下面一行中，我们记录上面一行数字是奇数还是偶数，每个奇数记为“1”，每个偶数记为“0”。下面一行中的数字序列构成了我们需要的二进制数。

十进制数	1	2	4	9	19	39	78	156	312
二进制数	1	0	0	1	1	1	0	0	0

求二进制数的除法表格

八进制与更高进制

很自然地，二进制数带来的一个问题是它们会形成由1和0组成的极长的数字序列，从而耗光宝贵的计算机内存。从理论上讲，我们可以采用希望使用的任意基，因此，缩短二进制数的一种方法就是用基为8的算术系统对它们进行转化。在这种“八进制”算术中，我们的字符集中需要有8种符号：0, 1, 2, 3, 4, 5, 6和7。

继八进制之后，下一个能够进一步将较大的二进制数压缩为简短表示的基是16，即采用十六进制系统。对于这种系统，一种便利的字符集选取方式是采用0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E和F这16个符号，其中A对应10，B对应11，C对应12，依次类推。要将一个二进制数转化为十六进制，只需将每四位划为一组（从右边开始），例如：

$$312_{10}=(0001)(0011)(1000)_2$$

再次参考刚才那个十进制-二进制转换表（忽略多余的0），我们得到结果：

$$312_{10}=138_{16}$$

不过，还有可能存在着完全由字母构成的十六进制数，例如，十进制数2748用十六进制表示为：

$$2748_{10}=10\times16^2+11\times16+12=ABC_{16}$$

关于十六进制的故事还没有结束。计算机的本领还在不断发展，而现在我们又开始熟悉以32和64为基的系统了。

数的漫长旅程

从最初刻在骨头上的标记到十六进制乃至更多发展，数所经历的这段旅程可谓漫长。在旅途的第一阶段，前进动力是人类社会的实际需求。但古人所关心的又不仅仅是数的用处。大约在公元前500年时，希腊古典时期[①]的公民们在他们称为逻辑的公共平台上使用自己的数系，诸如毕达哥拉斯和柏拉图等思想家已经致力于建立算术知识，即数论。数学独立的魅力和美从古代便已显现无遗。

2500年后，印度-阿拉伯十进制数系已证明其最为持久之特性，并为满足整个世界的数学、科学和日常需求而广泛传播。但数的历史还揭示出，在不同时代与不同地域中，各个文明是如何坚持用自己的方法处理数量和排序问题的。我们所拥有的数学遗产，包括数学的语言，见证了过去所施与我们的恩惠（巴比伦人的六十进制、罗马数系及二十进制）都对我们今天的数字词汇表有贡献。这是一笔丰厚而又多元化的财富。

---

① 古希腊历史中的一个特定阶段，大致上包含公元前5世纪到公元前4世纪的大部分，柏拉图和亚里士多德都活跃在这个时期。从时间上讲，它之前称为“原始时期”，其后则是“希腊化时期”。——译者注

3 为什么说质数是数学世界中的原子？
——建筑的砖石与算术基本定理

质数是埋藏在1, 2, 3, 4, 5 …计数数字中的珍宝。它们是构建整数数系的砖石，自古以来，一代又一代的数学家受到吸引，投身于寻找质数并且探索其玄妙性质的努力中来。大家对质数感兴趣的原因在于，它们是如此基本，却又蕴含着数学中最难战胜的未解之谜。

那么，什么是质数呢？首先，我们熟悉整数1, 2, 3, 4, 5, 6 …，它们沿着概念性的“数直线”而延伸。它们是数学中用来计数的数字，而埋藏于其中的“钻石”也就是我们所称的质数或素数。质数是不能被1和自身以外的任何数整除的整数。比如5是质数，因为用2、3和4去除它都会产生余数。但6不是质数，因为它可以被2整除而得到准确结果。式子$6=2\times3$表明它可以由更小的数字来构成，因此数学家使用合数这个词来称呼一个不是质数的数。合数可以拆分成较小的“因数”，而像5那样的质数则是不可拆分的。它是不可再分的，如同人们一度认为原子所具有的性质那样。

前几个质数包括2, 3, 5, 7, 11, 13, 17, 19, 23, 29…，但是在计数数字中，质数并不呈现为严格有序的形式。它们是数学世界中的“怪人”，看起来具有一定“随机性”。不过，这并不能阻止数学家去探寻其中的模式并梳理出可能发现到的一切规律。

在继续下面的内容之前，还有一些关于合法性的事情要理清，这涉及数1的问题。它是质数吗？由于不能被自身以外的任何数整除，它无疑是符合质数定义的。事实上，很多知名数学家，包括伟大的欧拉，都曾主张将1划入质数，不过目前的看法还是将数2作为第一个质数。（如果接受1是质数，那么我们经常用到的某些陈述就不再能应用于全体质数了。）①

即使我们把数2作为第一个质数，它还是具有特殊性质。它是所有质数中仅有的一个偶数：所有偶数都能被2整除，因此大于2的偶数都不可能是质数。于是除了2以外，在寻找质数时我们只需考虑奇数即可。此外，在数字位数超过1的情况下，以0或5结尾的数不可能是质数，因为以这两个数字结尾的数能被5整除；因此超过10的质数必定是以1, 3, 7或9结尾。但这并不意味着以这4个数字结尾的所有数都是质数：例如，在1到100之间以“1”结尾的数中，11, 31, 41, 61及71是质数，而21, 51, 81及91则不是。

质数的基本性质

关于质数，我们已经取得了大量成果。数学中最古老的分支——“数论”中充满了已证明的定理和对未来的挑战，而且每一年都会有数以百计的新定理出现在数学文献之中。重要的结果和伟大的思想以其简明和深刻诠释着数学发展的前景。

一个极为重要的信条是：任何整数都能表示为质数的乘积。例如，我们可以将41 184这个数表示为若干质数的乘积，即$2\times2\times2\times2\times2\times3\times3\times11\times13$，使用“幂”的符号，可以简写为$41\,184=2^5\times3^2\times11^1\times13^1$。因此我们说质数是构建整数数系的砖石。与此相关的一条性质是：将一个整数表示为质数乘积形式的方法有且仅有一种，即不可能

① 按今天的习惯，1既不是质数，也不是合数。——译者注

用几组不同的质数乘出相同的得数。这个关于唯一性的结论，称为算术基本定理。

从算术基本定理可以知道，任意整数都可以被至少一个质数整除。至少存在一个，但通常是有更多个，例如我们刚才提到的例子41 184，它可以被质数2、3、11和13整除。

> “我们是左右对称、有性别区分的两足生物，位于银河系的一条外旋臂之中，有能力识别质数……”
>
> 引自美国航空航天局外层空间探测器表面的文字

一个数何时为质数？

对于一个数值较小的数，容易确定它是否为质数。比如给定数N，我们可以进行快速的检验，通过观察还可能会节省大量的劳动。若数N为合数，也就是说$N=a\times b$，那么因数a和b不可能同时大于N的平方根。如果是那样的话，它们乘起来就会得到一个大于N的数。因此我们可以假设某一个因数（比如a）小于N的平方根。根据算术基本定理，我们还知道a必定可以被一个质数整除。将上面的结果放在一起即可发现，要判断N是不是一个合数，实际上只需考察是否有小于或等于N的平方根的质数能够整除N。若没有这样的质数，则N一定是质数。

例如，当我们研究211这个数时，首先必须计算它的平方根，结果大约为14.5。于是，我们只需考察211能否被小于14的质数整除，也就是说，要逐个检验2, 3, 5, 7, 11及13中有哪个能够恰好整除211。在计算器上完成这些计算，可以看到都会产生余数，由此我们断定211是一个质数。

很明显，这种方法对于有数千位数字的大数来说是不实用的，要检验一个很大的数究竟是质数还是合数是一个很有挑战性的问题。如何通过一次简单的检测而做出判断，人们几乎是无计可施。但是在2002年，一项重大突破出现了。3位印度数学家（Manindra

Agrawal、Neeraj Kayal和Nitin Saxena）证明：存在一种有效的算法或方法，能够检测任意数是否为质数（参见第20章）。

质数终结于何处？

质数有无限多个。我们可以说，存在着无限多个质数，也就是说，无论写出哪个质数，在它后面总还会有下一个。欧几里得在《几何原本》（第九卷，第20个命题）中给出了陈述这一事实的惊人定理，即“质数的数目比任意给定的数目更多”。其证明中最实质性的要点，就是今天我们所称的“欧几里得数”。

将到一个给定质数（我们称其为P）为止的所有质数乘起来，再加上1，就得到一个欧几里得数。用数学符号来表示，一个欧几里得数N具有如下形式：$N=(2\times3\times5\times7\times\cdots\times P)+1$。

让我们来看看$P=13$的情况，可以看到此时欧几里得数N不是质数而是合数：

$$N=(2\times3\times5\times7\times11\times13)+1=30\ 031=59\times509。$$

实际上，是质数的欧几里得数非常稀少，我们不得不等到$P=31$才能看到下一个出现，此时质数$N=2\ 000\ 560\ 490\ 131$。在此之后我们就要等到$P=379$，这时质数N大约有200位数字。而在$P=379$之后，随后出现的是质数的欧几里得数要到$P=1019$和$P=1021$时，1019和1021是一对毗邻的质数（称为“孪生质数”）。写出这些巨大的欧几里得数，就不是这里要讨论的问题了。

质数P	表　达　式	欧几里得数	是否质数
2	2+1	3	✓
3	(2×3)+1	7	✓
5	(2×3×5)+1	31	✓
7	(2×3×5×7)+1	211	✓
11	(2×3×5×7×11)+1	2 311	✓

P取较小值时的欧几里得数

现在我们可以欣赏欧几里得在关于质数有无限多个的断言中所显示出的才华。证明思路如下：首先任选一个质数P。我们的目标是证明存在一个超过P的质数。如同我们已经看到的那样，欧几里得数$N=(2\times3\times5\times7\times\cdots\times P)+1$或者是质数，或者不是——对于未特别加以说明的$P$值，我们不知道它究竟是不是。如果$N$是质数，那么我们就得到了超过$P$的质数，因为$N$无疑是比$P$大的。另一方面，就算$N$不是质数我们也没有失败。用$2, 3, 5, 7, \cdots, P$中的任一个质数去除$N$，都会得到余数1，因此它们都不能整除$N$。但是根据算术基本定理，我们知道，$N$必定可以被某个质数整除，由于这个肯定存在的质数不是$2, 3, 5, 7, \cdots, P$中的任一个，因此它一定比P大。这就是我们要找的比P大的质数。

于是我们证明，对于任意质数P总是存在着一个比它更大的质数，因此质数的数目必定是无限多个。如果质数的数目是有限的，那么我们就会拥有一个完全不同但也会因此而非常乏味的数学分支。

特殊的质数

在质数的无限集合中，有一些质数具有固定的形式。位于质数这门知识之核心地位的就是梅森数，它的名字来源于16世纪的数学家梅森。

梅森数具有$M_n=2^n-1$形式的表达式（其中符号n代表任意整数）。它们全都是奇数，因为从偶数2^n（它表示将数2自乘n次）中减去1必定会如此。尽管它们都是奇数，但却不一定是质数。要使M_n为质数，首先要求n必须也是质数，但这还不足以保证M_n是质数。尽管$M_2=3$, $M_3=7$, $M_5=31$, $M_7=127$都是质数，但梅森数M_{11}就不是质数，16世纪时人们就发现：

$$M_{11} = 2^{11} - 1 = 2047 = 23 \times 89$$

看起来梅森数不具有一般性的模式。M_{13}，M_{17}及M_{19}都是质数，但M_{23}不是，因为：

$$M_{23} = 2^{23} - 1 = 8\ 388\ 607 = 47 \times 178\ 481$$

事实上，大多数梅森数都不是质数，到目前为止我们只发现了47个是质数的梅森数。1963年，伊利诺伊大学赢得了发现$M_{11\,213}$是质数的荣誉。从那以后，人们陆续发现了一些更大的梅森质数，包括$M_{43\,112\,609}$，它是一个长达1300万位数字的质数。

在传统数学中，梅森质数可以用来构造完全数，即与自身因数之和相等的数。数6是一个完全数，因为$6 = 1 + 2 + 3$，28是下一个完全数，因为$28 = 1 + 2 + 4 + 7 + 14$。将欧拉与欧几里得的成果结合起来，数学家得到了一个利用质数（p）产生完全数的公式：当M_p为质数时，$2^{p-1} \times M_p$是一个完全数。

质数的频率

但是，在任意给定区间内，质数的分布情况如何呢？当我们在整数中前行时，质数会逐渐变得越来越稀疏。在数1~100之间有四分之一是质数，但在1~1000之间只有16%的数是质数，而在1~10 000之间质数所占比例就只有12%了。最近人们证实，在从$1 \sim n = 10^{23}$的范围内有1 925 320 391 606 803 968 923个质数，这意味着在这个巨大区间之内实际上只有占总体1.9%的数是质数。

活跃在19世纪早期的伟大数学家高斯早在16岁时就沉迷于质数表之中，并且着手用图表方法来研究质数在不同区间内所占的比例。这个比例，即通常所说的质数的“密度”，是用下列方式计算出来的：

$$\frac{\text{小于}n\text{的质数的总数}}{\text{小于}n\text{的数的总数}}$$

通过坚持不懈的计数，他推测出了一个能进行理论估计的公式。

数（n）	小于n的质数的实际个数	实际密度	理论估计
10	4	40.0%	43.4%
100	25	25.0%	21.7%
1000	168	16.8%	14.5%
10 000	1229	12.3%	10.9%
100 000	9 592	9.6%	8.7%
1 000 000	78 498	7.8%	7.3%
10 000 000	664 579	6.6%	6.2%
100 000 000	5 761 455	5.8%	5.4%
1 000 000 000	50 847 534	5.1%	4.8%

质数的实际密度与理论密度

高斯的公式（以n的对数的形式给出）来自于实验验证，而对该公式的数学证明在19世纪初期还是遥不可及的。到19世纪末时，对理论估计公式的证明出现了，但终究还让人觉得不是很满意，因为证明依赖于微分学和积分学（参见第8章），而将这些技术引入纯粹的“算术”被认为是过于人为的。这样的批评在50年后消失了，那时出现了一个纯粹的算术证明——尽管其论述极端之晦涩而冗长。

这个已经证明的定理在今天被称为质数定理（prime number theorem，PNT），它的内容是：通过选取足够大的n，可以使质数的实际密度与其理论估计之间的差异减小到我们需要的任意程度。数学家们将PNT看做是这门学问中最伟大的定理之一，特别是由于它还与大家心目中最伟大的英雄之一（高斯）联系在一起。而高斯的学生黎曼以其1859年那篇著名的论文指出，质数的长程分布是怎样依赖于“黎曼zeta函数”的性质以及“黎曼假设”的。

对规律性的探寻

幸亏欧几里得，质数的无限性才为数学家们提供了一个广阔无垠的舞台。各种各样的问题由此而产生，比如“质数的外观形式是否有什么规律性？”或是“我们能否得到只生成质数的公式？”

18世纪时，欧拉给出了一个著名的公式——n^2+n+41（其中n还是代表任意整数）。它的特殊之处在于，当我们将$n=1, 2, 3, \cdots, 39$代入公式时，所得结果都是质数。比如选取$n=7$来加以考察，我们得到$7^2+7+41=97$，它确实是一个质数。事实上，一直到$n=40$时情况才发生变化，此时得数为41×41，不再是质数。

多变量表达式可以产生更好的结果，最近有人指出，存在一个公式（尽管相当复杂）能够产生所有质数。麻烦在于它包含10个变量，而且某些变量的指数可能会非常大（具有10^{45}的数量级）。此外，它也会产生不是质数的结果。

通过转而研究不同形式的表达式即可获得另一种类型的成果，让我们来考虑只能产生质数但不是所有质数的表达式。1947年，威廉·H. 米尔斯得到了一个引人注目的成果，他证明存在着一个数，用字母k来表示，它满足：

对于任意值n，不超过k^{3^n}的最大整数总是一个质数。

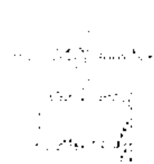
仅有的困难在于，没有人对k的值有多少了解，只知道它存在，而且可能不止一个。今天我们称它为米尔斯常数。人们认为，如果黎曼假设为真的话，k可能取得的最小值大约为1.306 377 883 8。这些我们非常渴望获得的关于质数分布的结果依赖于黎曼假设的真实性。实际上，有如此多的关于质数以及其他很多主题的结果依赖于黎曼假设，人们无比热切地等待着它的证明，而它可能仍旧是整个数学领域中最具挑战性的难题（参见第20章）。

质数的首要地位

在对于质数的讨论中，我们进入了数论的领域，这个领域一度曾被断言是数学中没有任何实际意义的一个分支。这个断言可能曾经是事实，但在今天，数论在当代编码科学和计算机安全领域中处于核心地位（见第14章）。计算机安全依赖于这门深奥的学问，在这里，欧几里得、梅森、欧拉、17世纪的大法官费马以及其他很多学者的定理获得了新的生机。

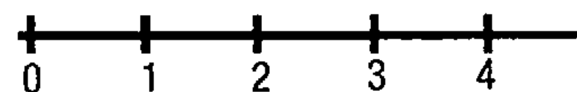
质数的定义方式很简单，但即使已为我们所熟悉，它们仍然以种种艰深费解而又引人入胜的方式向我们提出持续的挑战。它们吸引了世界上最出色的数学家的关注，无可争议地说，它们是构成算术大厦的砖石，并且处于如此众多之数学成就的中心地位。将它们称为数学中的“原子”是毫不夸张的，追查它们的未解奥秘始终是数学中的根本追求。

4 最奇怪的数是谁?
——实数、无理数与超越数



代文明为今天的计数系统奠定了基础，开始于1,2,3的整数序列，无论采用哪种符号系统，已被证实是价值无可估量的实用工具。但是随着时间推移，这些已经不足以满足我们的数学需求。人们需要新型的数，凑不足整数的数（分数）和小于零的数（负数）的概念应运而生。建立数系的努力并未到此而止，甚至产生出一些实际上非常神奇的数。

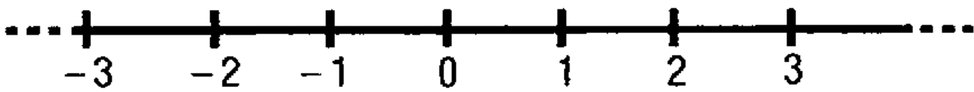
小孩子要学习的计数数都是整数（1, 2, 3等）。为了强调它们作为绝大多数研究的起点地位，数学家们喜欢称之为“自然”数，他们还喜欢把它们叫做正整数，“整数”表示完整而没有零头的数。（如果把零也考虑进来，则称这个数系为非负整数。）从概念上讲，考虑一下这些沿着数直线排列的构建数学大厦之砖石会是有所帮助的。



数直线上的零和正整数

但是，仅就其自身而言，对于很多数学计算来说这是远远不够的。首先，我们需要在数直线上增加负数：−1, −2, −3, −4, −5等。商业活动是我们可能遭遇负数的最清晰的情境之一。比如说，银行账单上出现的令人沮丧的数字“−250”表示我们已经透支了，因而亏欠着银行这么多钱。负数还可以用来度量零度以下的温度。负数的绝对值越大就表示越寒冷，例如−20℃就比−10℃更冷。这

种给负数排序的方法是简明易懂的，可以写为 $-20<-10$。将负数引入数直线，则随着绝对值增大，负数（这里是指负整数）从右向左延伸。将正整数、负整数和零合起来，就得到数学家所说的整数集。



数直线上的整数集（包括负整数、零和正整数）

今天，我们可以非常愉快地接受负数的存在，但是在过去，负数的反对者对于其存在都是怀疑的。18世纪时，一位对数学略有涉猎的英国外交官马塞雷认为负数是荒谬而毫无意义的。他声称它们“玷污了所有关于方程的信条”。这当然是无稽之谈，不过后来者可以看到，他这番厥词恰好使负数深深地植根于数学土壤之中。

从整数到分数

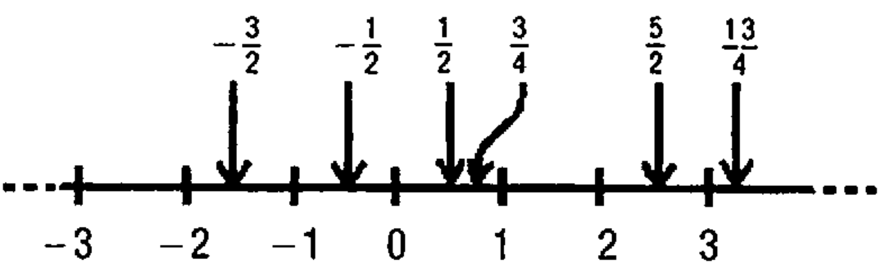
但是负整数还不能解决如何表示整数的一部分的问题，这使得分数应运而生。我们经常把分数看做是单位量的一部分。比如有20个人参加一个生日宴会，而生日蛋糕只有一个，那么在理论上，每位客人将会分得二十分之一或 $\frac{1}{20}$ 个蛋糕。上面的数叫做分子，下面的数叫做分母。数学家们经常将分数称为有理数[①]，这并不是因为它们明白事理，而是因为它们具有一个数除以另一个数的表示形式，或者说表示为比例的形式。

> 整数是一切数学的根源。
>
> 闵科夫斯基《丢番图逼近》(1907)

① “有理数”原文为“rational number”，其中rational这个词来源于ratio，是“比例”的意思。在近代数学知识引入中国时，一度曾将“rational number”颇为恰当地译为“比例数”，可惜的是后来错误译法反而占了上风，其后以讹传讹，日益盛行，成为今日之遗憾。——译者注

像这种类型的分数，将1放在任意非零整数（用代数方式将其记为n）的头上，我们称之为单位分数。古埃及人发明了一种用单位分数进行计算的有趣方法，由此引发的一些数学难题直至今日仍未完全解决。不过今天我们所用的正常分数并不将分子限定为1。比如，当参加生日宴会的人有两个相同的蛋糕可以分享时，就可以每人分得$\frac{2}{20}$个蛋糕。有时可以利用除法将分数化简，比如将$\frac{2}{20}$的分子、分母同除以2，就能得到一个等价的结果$\frac{1}{10}$，它表示每位客人能够分得一个生日蛋糕的$\frac{1}{10}$。对分数进行这样的“约分”，是一种有用的工具，也是处理分数的一种重要技巧。

回过来再看数直线这种可视工具，我们也能将分数表示在它上面：只需在整数之间合乎比例的位置上将分数插入即可，例如$\frac{1}{2}$位于0和1之间的中点处，而像$\frac{5}{2}$这样“头重脚轻”的分数则位于2和3之间的中点。



数直线上的分数

数直线也引起了新问题。我们能够继续在上面添加分数吗？总共能够挤进去多少个分数呢？实际上，总是可以有更多的空间，例如，假定我们已经将$\frac{2}{7}$和$\frac{5}{17}$插入数直线，那么分数$\frac{7}{24}$（分别将两个分数的分子和分母相加即可得到）就会位于两者之间。你可能忍不住要说，分数填满了整个数直线，不再留下任何空隙。

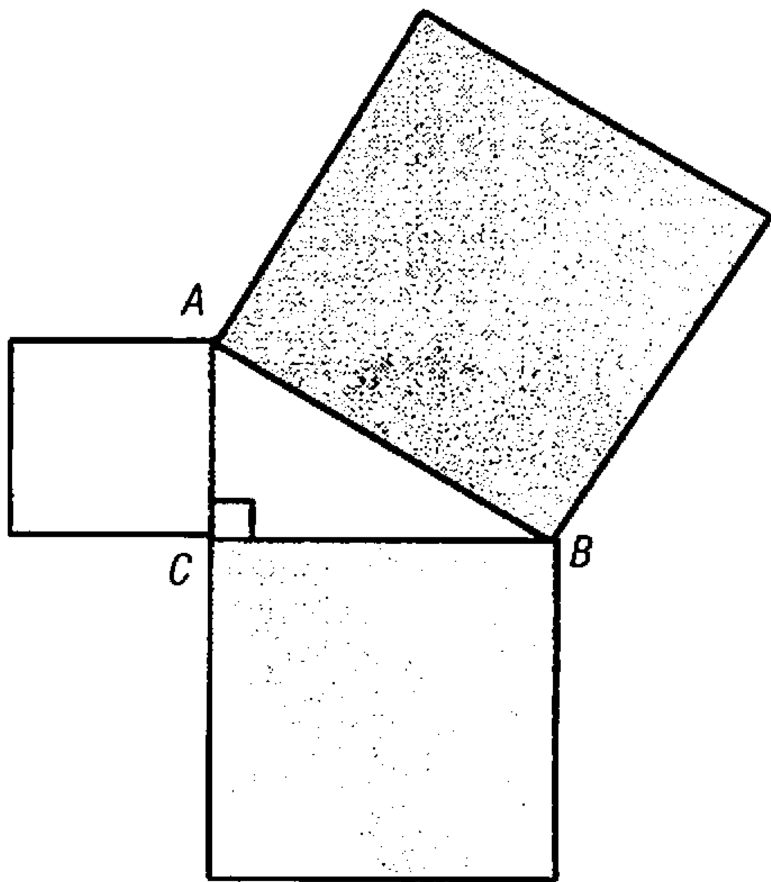
分数与毕达哥拉斯学派

大约公元前550年时，在毕达哥拉斯领导下的一群希腊学者将数

看做是理解自然的关键。就某种意义而言，他们的观念是类似于神秘主义的；数被赋予了人类的属性，偶数被看做是女性的，而奇数则是男性的。但是古希腊人也有一套清晰有力的处理数的方法，这体现在他们那特别关注分数的度量理论中。

在毕达哥拉斯学派的度量理论中，长度是通过用一个给定的较小单位去分割来度量的。这个单位可以是1英寸、1毫米或者是别的什么——单位的名称对这种理论不会有什么影响。将所选单位置于待测长度的起点处，接着首尾相连地连续放置——就像用步长方法测距离时，脚趾挨着脚后跟一步一步直线前行那样。如果我们足够幸运，该长度可能由整数个单位构成，比如说14个。但是更有可能的是，最后一个放置的单位并未与长度的末端对齐，在这种情况下，毕达哥拉斯学派的几何学家会添上另一个长度，它与第一个长度相同(还可以继续添加)，再继续度量直到最终获得“对齐”的结果为止。

接着，他需要通过除法来获得一个长度的度量结果。比如说，当27个单位恰好与2个长度的末端对齐时，单独一个长度的度量结果就是分数$\frac{27}{2}$，也就是$13\frac{1}{2}$个单位。因此，毕达哥拉斯学派度量理论的特别之处，就是它总是产生表示为分数形式的结果。

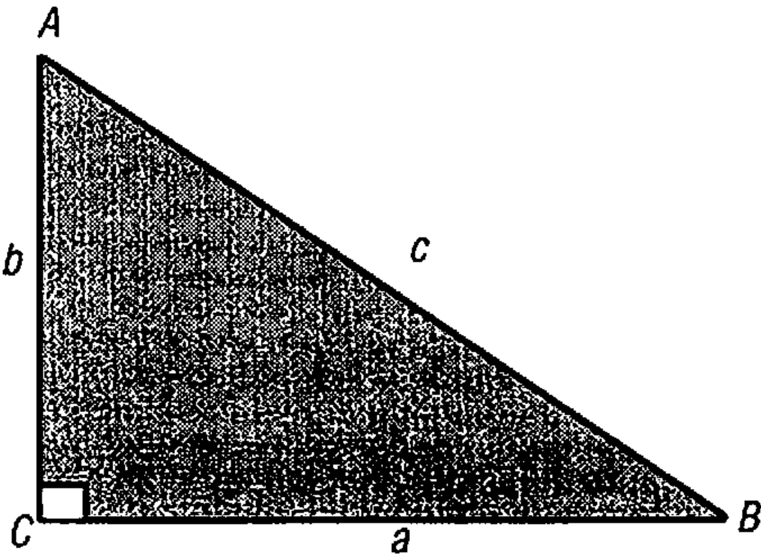


直角三角形边长上的正方形

分数无能为力的情况

在数学中，使毕达哥拉斯获得巨大声望的是他那关于三角形的不朽定理：直角三角形斜边（即直角三角形中与直角相对的边）上正方形的面积

等于其余两边上正方形的面积之和。这是一种会对实际度量（以及分数）产生影响的理论，因为古希腊人就是使用正方形这样的词汇来思考几何问题的。所以，毕达哥拉斯定理是用附在边上的真实的正方形图形来表示的。如果我们用a, b, c来表示三边之长，就可以得到这个定理更为人所熟知的形式：等式$a^2+b^2=c^2$。



直角三角形的毕达哥拉斯定理

但是我们不妨设想一下，当a和b的长度都是1时，也就是$1^2+1^2=c^2$，于是$c^2=2$。因此斜边的长度必定是2的平方根（记作$c=\sqrt{2}$），近似为1.414 2，因为1.414 2 × 1.414 2得到1.999 961 64。结果接近于2但是并不等于2。

令人惊诧的事情发生了。按照毕达哥拉斯学派的度量方法，如果想用三角形边长$a=1$作为单位，通过首尾相连的方式去度量边长c（它应该等于2的平方根），我们就会发现，无论将多少个长度c连接起来，永远都不可能用单位与它们在末端对齐。这意味着定义为$\sqrt{2}$的长度不可能是一个分数。根据主流的度量理论，一切长度都应该具有分数形式。但是$\sqrt{2}$显然与之矛盾。

毕达哥拉斯学派无法应对这一惊人的新发现。根据传说，学派中的一位成员（希帕索斯）因为泄露了$\sqrt{2}$的秘密而被抛入大海。他们将自己的怒火发泄到泄密者身上，而不是对其进行奖励。然而，就算是住在奥林匹斯山的诸神也无法改变关于$\sqrt{2}$的尴尬现实。存在着一个可以实际构造出来的长度，（还有比构造一个具有两条相等边长的直角三角形更容易的吗？[①]）但是它的第三条边却无法用现有理论来度量。毕达哥拉斯可能会说，如果我们再坚持一下，放置更

① 在欧几里得《几何原本》中，第一卷第46个命题可以实现这一目的。——译者注

多的单位就可以获得具有分数形式的结果，并以此来摆脱困境。但是关于$\sqrt{2}$不是分数因而不可度量的逻辑证明的出现，将他的理论彻底摧毁。因此，毕达哥拉斯没有任何办法来应对这一被揭露的事实。

前方还有更多不受欢迎的新发现呢。继2的平方根不可能表示为分数形式的发现之后，一系列其他的平方根也呈现出类似的矛盾性：$\sqrt{3}, \sqrt{5}, \sqrt{6}, \sqrt{7}, \sqrt{8}, \sqrt{10}$等——不出现于其中的只有诸如4, 9, 16, 25等完全平方数的平方根。于是我们发现，除了整数和分数（即有理数），我们还可以在数直线中插入无限多个数学家们称之为*无理数*[①]的数，它们潜伏于分数之间。

有理数加上无理数就构成了数学家们所说的“实数线”。任意实数都可以置于直线上的某个位置。但是用消极方式来定义事物（即将无理数定义为不是有理数的数）并没有充分揭示其本质。戴德金和康托尔在19世纪后半叶对这一主题进行了探讨，并且各自得到了一种在数学中具有持久地位的有用定义。戴德金将一个实数定义为一个“戴德金分割”，于是，在将数直线划分成两部分时，一个“分割”就界定了一个实数；而康托尔则将实数定义为某种有理数序列的极限，例如$\sqrt{2}$就是有理数序列1, 1.4, 1.41, 1.414, 1.4142,…的极限。利用他们的成果，数学家得以在更高的精确水平上处理关于实数的问题。

从无理数到超越数

数学家为了确定2的平方根（以及其他一些数的平方根）的无理性而设计出来的证明，是一个精巧的证明。但是它对于无理数的本质并没有产生什么进一步的洞察。

还有另外一种贯穿有理数/无理数概念的分类，就是代数数。代

① 与前述“有理数”类似，实际上“irrational number”的较为恰当的意思应该是“非比例数”，这里只能遵循国内习惯来翻译。——译者注

数数是指能够作为某个只涉及x的幂的方程（例如$x^2-4x+1=0$）的解的数。

无理数$\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt{7}$, $\sqrt{8}$及$\sqrt{10}$等都是代数数，以$\sqrt{2}$为例加以说明，它是方程$x^2-2=0$的一个解（$\sqrt{3}$则是方程$x^2-3=0$的一个解，以此类推）。代数数的分类包含了有理数和某些无理数。

所有的实数都是代数数吗？回答是否定的，因为在代数数的边界之外还存在着一些“不羁的”无理数。这些数不满足任何代数方程，18世纪时瑞士天才数学家欧拉将它们描述为“超越的”。意思很简单，它们“超越了代数”。但是要说明一个数是超越的（换言之，它不是任意只涉及x的幂次的方程的解）可是一个与众不同的挑战，原因在于要证明一个否定的表述可是众所周知地困难。

超越数的分类中包括两个在数学中最久负盛名的数。最著名的当属“pi”，表示该数的符号来自于希腊字母π，它是圆的周长除以其直径所得的结果，数值约为（从它的无限小数形式中取前6位小数）3.141 592[①]。

继π之后，常数e享有超越数中排名第二的声望。这个数近似等于（取前6位小数）2.718 281，某些场合中也被称为“欧拉数”，这个名字来自于将它加以推广的那位数学家。我们知道，它在处理诸如人口增长和商务数学等问题中具有无可估量的重要性，而且在概率论和统计学中也占有一席之地。19世纪时，埃尔米特证明e是超越数。通过改进埃尔米特的方法，林德曼证明π也是超越数，从而解决了一个“古代谜题”。利用这一发现可以对一个著名的谜题给出结论：“化圆为方”，也就是说只用直尺和圆规来构造与给定圆形具有相同面积的正方形，确实是不可能的。

---

① 作者在截取前若干位小数时，并没有做舍入处理，下文中e的近似值也是如此。——译者注

与林德曼关于π的决定性研究大约同时，康托尔在数学界引起了一场轩然大波。当时数学家对于分数、有理数以及像e和π这样的少数几个超越数是熟悉的。康托尔证明几乎所有的“实”数（或者是有理数，或者是无理数）都是超越数，从而推动了对于传统数直线之本质的再思考。如果我们将数直线想象成只由分数、整数以及诸如$\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt{7}$, $\sqrt{8}$ 及 $\sqrt{10}$ 等数所组成，那就大错特错了。与聚集在它们周围的超越数相比，这些数实际上分布得相当稀疏。

考虑到超越数远比其他各种类型的数要多很多这一新发现，使我们仍旧感到奇怪的是，除去π和e，我们实际上能够识别的超越数实在不多。如果愿意的话，我们可以进行一些儿戏式的操作，得到诸如7π, π^2+1, e^2及e^3+1等数，但这只不过是蜻蜓点水而已。而且我们必须小心。当我们开始随意地混合超越数时，就会遭遇困难。比如，我们甚至不知道$e+\pi$是不是超越数。数e的π次方是超越的，但是π的e次方是否也可以这样说呢？目前还不知道。

尽管如此，我们还是取得了一些进展。1900年，大数学家希尔伯特为新世纪的到来向他的同行们提出了著名的“23问题”的挑战。其中第7个问题就是说明$2^{\sqrt{2}}$(即2的$\sqrt{2}$次方）是超越的。30年后，盖尔丰德解决了这个问题，他还发现了整整一族的超越数。

十进小数的影响

数学家喜欢用e和π那样的字母来表示数，对于某些数还可以用平方根符号，而在实际应用中则需要十进小数。举个最简单的例子，1/4可以表示为0.25，3/5可以表示为0.6。当然，可以轻松转换为小数形式的分数并没有那么多。比如用十进小数表示1/3，就会得到无限重复形式0.333 33…但是我们可以断言，一个分数的小数展开形式具有规则性，即一种循环模式。例如，分数2/7用小数形式表示得到

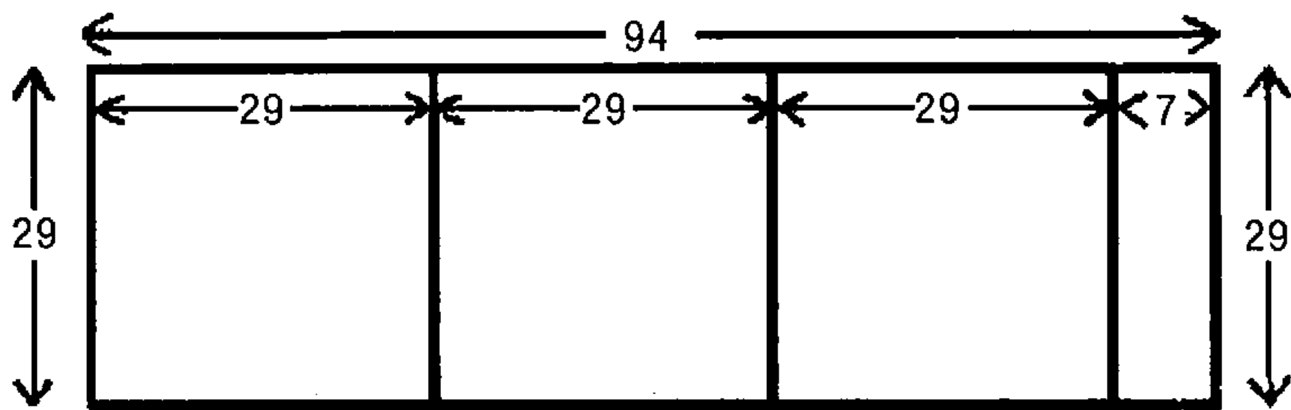
0.285 714 285 714 285⋯，它会一直重复285 714这个序列。

用小数表示无理数时也会得到无限形式，但是没有模式，不过这一分别仍不足以区分代数数和超越数。代数数 $\sqrt{2}$ 和超越数π都具有无模式的小数展开形式：

$$\sqrt{2} = 1.414\,213\,562\,373\,095\cdots$$
$$\pi = 3.141\,592\,653\,589\,793\cdots$$

如 DNA 般的连分数

考察数的连分数展开形式，是一个更为深刻的问题。可以把它看做是一种几何理论。让我们先来研究一下 $5\frac{29}{94}$ 的连分数展开。将注意力集中在分数部分 $\frac{29}{94}$，我们可以想象一个短边为29、长边为94的长方形。



94=3×29+7

29=4×7+1

7=7×1+0

数的连分数展开

我们最多能够从这个长方形中减去3个（边长为29的）正方形，余下一个短边为7、长边为29的长方形。如果我们对剩余部分重复前面的操作，会发现可以减去4个正方形（因为29＝4×7＋1），最终得到7个单位正方形。$5\frac{29}{94}$ 的连分数就是由我们刚才减去的正方形数

(3, 4, 7) 所决定的，它可以表示为：

$$[5; 3, 4, 7]$$

连分数就像是一个数的DNA一样。在显微镜下，$\sqrt{2}$ 具有规则的连分数表示：

$$\sqrt{2} = [1; 2, 2, 2, 2, 2, 2, 2, 2, \cdots]$$

这是因为，在任一步骤中我们都可以减去两个正方形，而且永远无法穷尽。$\sqrt{2}$ 是非常规则的，而超越数e的连分数展开形式则呈现出另一种不同的模式：

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, \cdots]$$

大名鼎鼎的π则始终还是一个谜。它具有高度不规则的连分数展开形式，这个桀骜不驯的数没有呈现出任何屈服的迹象：

$$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, \cdots]$$

扩展中的数的范围

当人们仅仅希望将自己面前的物理实体（牲畜、庄稼和土地）加以量化时，会产生一个感觉：正常的计数数就足够用了。但是若干个世纪以来，方便好用的概念性工具数直线带来了根本性的变化。负数的出现表明，数直线不仅可以从左向右增加，还能从零点从右向左减少。2的平方根给古希腊毕达哥拉斯学派带来的数学困境有力地动摇了令人感到舒适的分数观念，即数学家所谓的有理数。数的世界呈现出怪异和不确定的形式，从而拓展到今天所称的“无理数”，和更理论化但很少具体介绍的超越数。从有刻痕的长棍开始，延绵至今，这真是一条漫长的道路。

5 虚数真是虚幻的吗？
——从虚数"i"到八元数

漫主义诗人塞缪尔·泰勒·柯勒惠支笃信人类的想象力与个人体验的力量，对于数学他抱有一种谨慎的态度，在他看来这门学问"理性泛滥而想象力匮乏"。但是早在两个世纪以前，哲学家笛卡儿就已经在谈论"虚幻"之数的数学王国。尽管人们最初以极其谨慎的态度来看待虚数，但最终还是接受了它们，将其纳入数学知识大厦。

由牛顿和莱布尼兹所设计的、在微积分中用来计算变化率的方法呈现出一种新的形式，数学家们创造出他们所称的"复分析"作为处理几何学和数论问题的一种强有力的技术。由于引入"虚数"而产生的全新的代数系统挑战了人们对于代数学之本质的原有观念。而最不可思议的事情，大概要算"虚"数在实际问题中的突出作用。

那么，虚数究竟存在于何处呢？其中某些会比我们原先以为的更熟悉。我们可以利用简单的计数数，就最基本的情况对其思想进行简要的介绍。比如我们的果盘中有5个苹果，拿走4个，还剩下1个。要是将5个全都拿走，那就只剩下0个。但是我们不可能拿走6个苹果，因为开始时果盘中只有5个苹果。

在进行银行业务交易时，我们就会与一种令人苦恼的不同类型

的数打交道——欠债的数目。比如，我们的账户上可能有5英镑，但是我们还有透支能力。因此在提取7英镑后，就欠银行2英镑，这可以表示为-2英镑。当然，我们不可能实际上看到或者拿到-2英镑这个负的数量（在这个意义上讲，它并不存在），但是我们确实知道如何解释自己的债务，而且银行也是知道的。这意味着我们可以有效地用虚幻的-2英镑硬币进行计算，例如，在存入10英镑后，第二天我们就能知道银行会取走2英镑作为对债务的偿还，而账户上的结余将是8英镑。用数学记法来表示，其和为$10+(-2)=8$，这和说$10-2=8$是一样的意思。

从这个角度来看，负数也可以被认为是“虚幻的”，而且在最初遇到负数时人们也确实是这样认为的，正如我们在前面内容中（参见第4章）已看到的那样。现在它们看来不那么引人争议了，而将它们纳入数学体系则使我们有效地扩大了数字系统。在数学理论中，我们可以用方程的方式来将负数及其用处概念化。如果仅限于考虑正整数$1, 2, 3, \cdots$，我们就只能解类似$x-1=0$那样的方程，其解为$x=1$。但是我们没法解$x+1=0$那样的方程。将数$0, 1, 2, 3, \cdots$中的任意一个代入x都不可能得到等于0的结果。我们必须到正数以外的范围去寻找方程的解，因此，只有承认负数我们才能够得到方程的解，即$x=-1$。通过想象负数的概念，我们就在数学的意义上使它成为存在物。

二次方程与虚数“i”

在“二次方程”的背景下，虚数得到了进一步的发展，对于很多成年人来说，“二次方程”是一个自校园时代开始就一直回响在记忆中的短语。这是一类古老的问题，古埃及人在计算土地面积的问题中就曾解过二次方程。二次方程的定义性质是涉及一个未知数的平方，不妨用x^2来表示，用“quadratic”这个词来描述它，来源于有

四条边的正方形这个概念[1]。在实际应用中，与二次方程相关的数学知识不限于任何一类特殊问题。

但是，在类似$x^2+1=0$这样的二次方程中会发生什么呢？如果用常规方法来处理这个方程，就需要x^2等于-1才能使等式左边为0。但这是不可能的，因为x^2不可能等于一个负数。（比如，当$x=5$时，$x^2=25$；而当$x=-5$时，x^2还是等于25！）

克服这个困难需要想象力的又一次飞跃。这次的问题是，我们要给$x^2+1=0$发明一个解，并且给它起一个名字，也就是数学家所称的“i”，它代表虚幻的。由于我们已经宣称它是一个解，因此有$i^2+1=0$，或者换个说法就是$i^2=-1$。17世纪时，笛卡儿在这一数学背景下最早使用了“虚幻的”这个词。他认为所有的二次方程都应该有解，但是他承认“在很多情况下并不存在与我们所想象的相对应的那种量”。

关于i这个符号，我们要宣称的一切就是$i^2=-1$，而这就是我们需要知道的一切。我们不需要花时间去讨论关于i的本质的问题。在这个层面上进一步深究（如同某些人曾经做过的那样）以及将i描述为-1的平方根，都是错误的方向。因为那会暗示着它是作为实数而存在于数直线上的，就好像我们能够在计算器中输入-1，再按下$\sqrt{\ }$键而得到一个实数结果。

这个符号使我们能够对一系列没有实数解的二次方程求解。例如方程$(x-1)^2=-4$，可以将它改写成：

$$\left(\frac{x-1}{2}\right)^2=-1$$

① 正方形的英文写法是“quadrate”，它是古代数学中用来表示平方概念的几何形式。——译者注

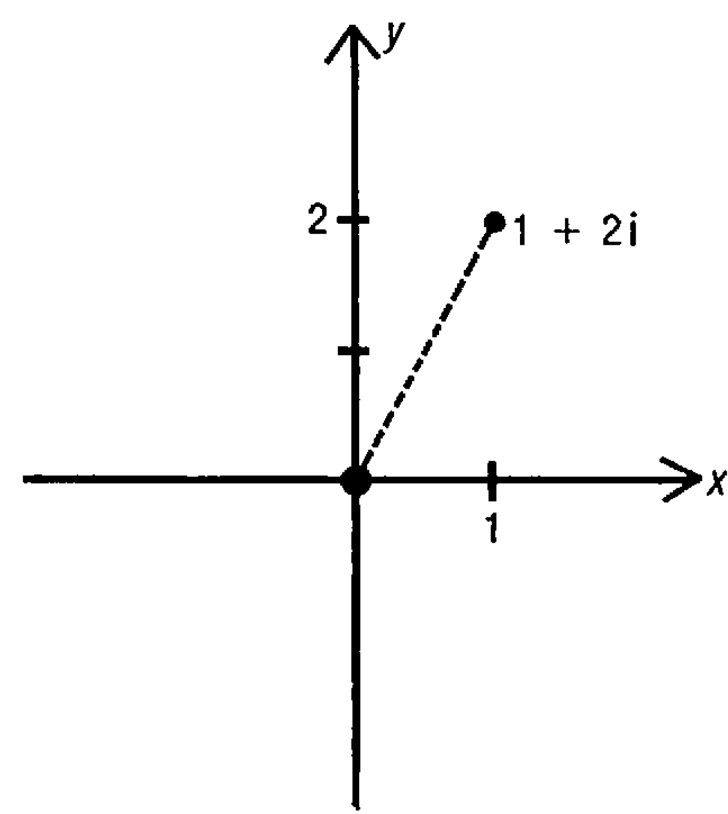
由此可知，括号内的表达式就等于i，从而得到解$x=1+2i$。实际上，这个方程有两个解，另一个是$x=1+(-2)i$。

复数

数学家们将我们刚才所得到的那种类型（比如$x=1+2i$）的解称为复数。推而广之，当a和b表示实数时，表达式$a+bi$具有“复合的”含义。不过到了这个阶段，有人可能会怀疑，我们是否已进入一个只为探讨不切实际的问题而发明出来的充斥着虚构的解的领域，这种疑虑是可以理解的。当人们认识到复数拥有具体的几何含义后，围绕在它们周围的神秘气氛才部分地消散开来。对这一成果作出重要贡献的，当属爱尔兰神童威廉·卢云·哈密顿，他在19世纪30年代的著述推动了将复数$a+bi$看做有序数对(a, b)而完全抛开i来进行处理的思想。

于是，数学家们有了两种书写复数的方式。周末他们想改变对待复数的方式时就写作(a, b)，而在工作日则将它们写作$a+bi$，视具体情况而定。

如同我们已经看到的那样（参见第3章），实数位于一维数直线上，向右（正数）和向左（负数）延伸。而复数则是二维的。一个复数可以表示为平面上的一个点，这时将所得到的图像称为阿尔冈图，这个名称来自于19世纪早期的一位先驱，法国业余数学家阿尔冈。



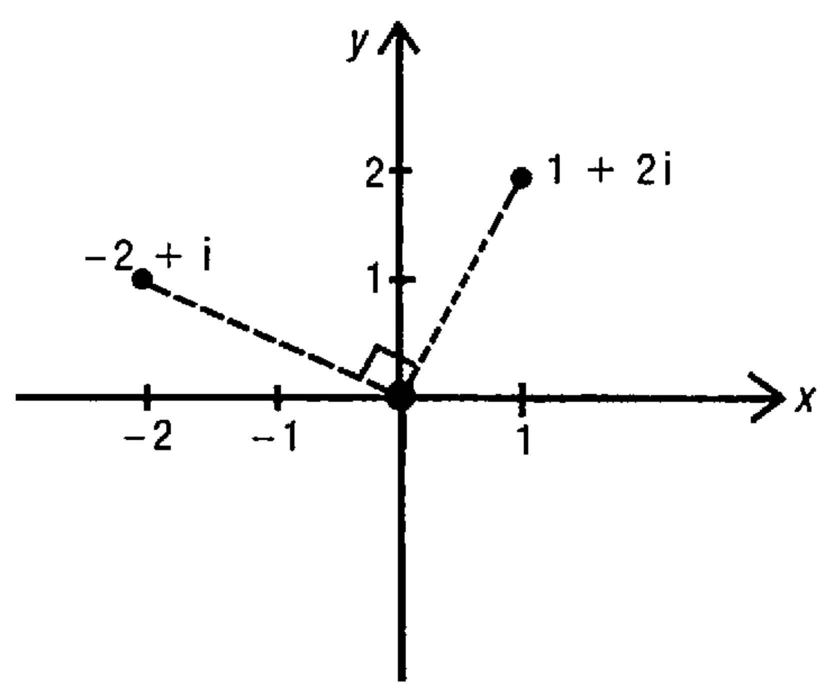
复数$1+2i$，二维平面上的一个点

以复数$1+2i$为例，我们可以赋予它一种图像表示，即横坐标x等于1而纵坐标y等于2的点。按照坐标位置的书写惯例，显然应该将这个点记为$(1, 2)$。

不过，i的几何应用可并没到此为止，因为它还有更进一步的含义。将任意一个复数乘以i，就会得到与原复数呈垂直关系的另一个复数。例如，还是用$1+2i$这个例子：

$$i\times(1+2i)=i+2i^2=i-2=-2+i$$

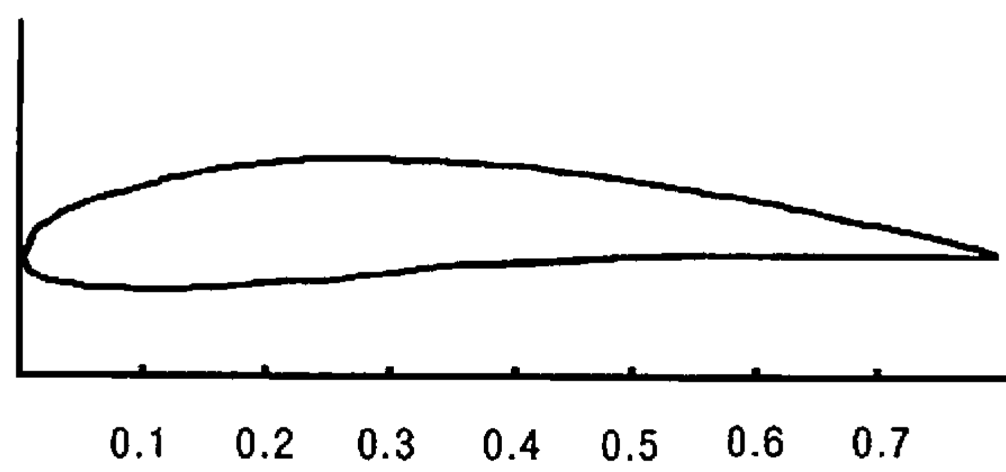
因此，乘以i的结果就是将平面上的点(1, 2)旋转到点(−2, 1)，如图所示。



复数$-2+i$是$1+2i$旋转的结果

于是符号i就产生了两种几何含义，平面上的一个点，或者是转过一个直角。

与数直线上的数一样，复数也可以进行加、减、乘、除运算。它们是对我们原有数系的进一步丰富。一个全新的数学分支（复分析）从复数的世界中产生，人们发现，相比起以实数为基础的对应理论，实际上它更为精巧而广阔。它具有微积分的形式（参见第8章），其变量是复数，而此前的微积分中变量都是实数。这种新的理论与包括几何学在内的数学中其他领域具有千丝万缕的关联，19世纪，黎曼将它应用于自己对于质数分布的理论研究之中。



茹科夫斯基机翼的阿尔冈图

复分析还有大量的实际应用。飞机设计就是其中之一。例如，我们取z为阿尔冈图中位于半径为1的圆周上的点，并计算$z+\frac{1}{z}$的值，就可以描画出茹科夫斯基机翼的轮廓。

复分析在现实世界中最著名也是不可或缺的应用，当属对电学和电子工程学的研究，不过在这两个领域使用的符号是*j*而不是i，这是为了避免与专门用来表示电流的符号i相混淆。

从复数到四元数

在对复数理论做出了重要贡献之后，哈密顿继续思考，是否可以将复数进一步扩充而使其系统更为丰富。数学家已经将数的概念从计数数扩展到了负数，又从实数扩展到了复数。他们还能更进一步扩展吗？

正如哈密顿所看到的那样，前进方向就是将二维的复数扩展为三维的数。历经多年思考而一无所获之后，他终于在1843年捕捉到了灵感的火花。原来他一直都走错了路：不应该去搜寻三维的数，而应该试试四维的。结果就是哈密顿的“四元数”。

四元数可以用符号表示为$a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k}$的形式，这与复数的$a+b\mathrm{i}$形式是类似的。如同复数中的$\mathrm{i}^2=-1$那样，四元数中的“虚数”i、j和k也满足$\mathrm{i}^2=-1$，$\mathrm{j}^2=-1$及$\mathrm{k}^2=-1$。

为了便于记忆四元数乘法规则，可以按几何解释将符号i、j和k看做是一个圆周上的点，每一个字母乘以下一个字母刚好得到第3个字母。但是要记住，逆时针相乘会得到正的结果，而顺时针相乘则会得到负的结果，因此$\mathrm{i}\times\mathrm{j}=\mathrm{k}$（逆时针），而$\mathrm{j}\times\mathrm{i}=-\mathrm{k}$（顺时针）。

让很多数学家也觉得难以接受的事情是，乘法的顺序也是个问题。在常见的算术问题中，无论是7乘5还是5乘7都会得到一样的结果，对于二维的复数来说情况也是如此。不过，四元数中的新奇之处并没把哈密顿搞糊涂，事实上它是开创性的。由于有了四元数，一种新类型的代数诞生了。

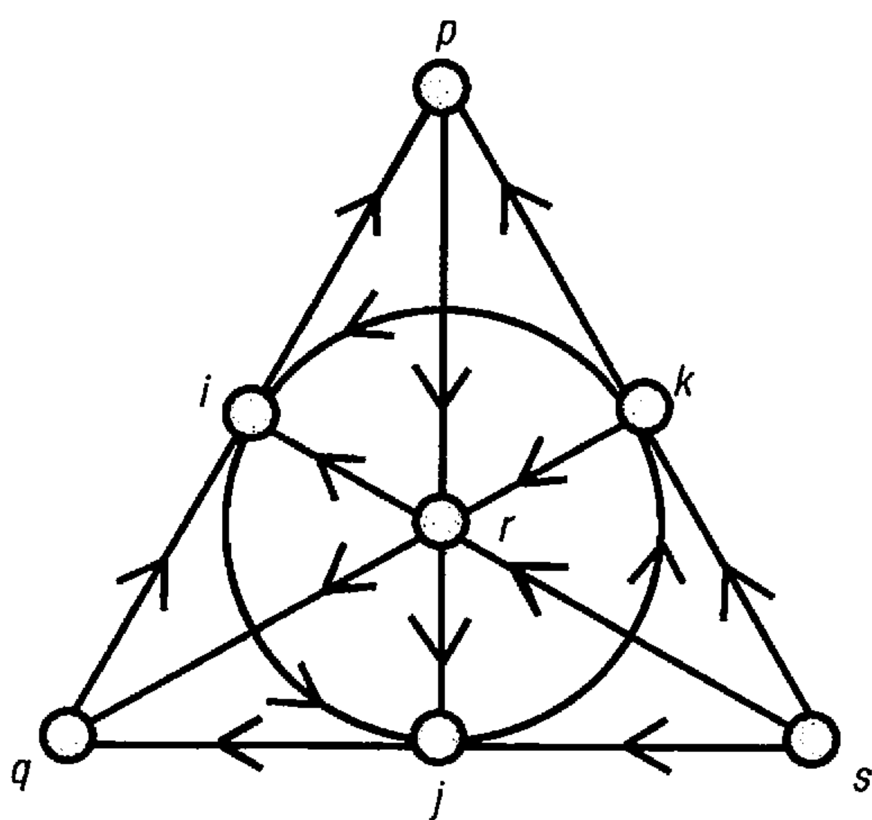
还有一个问题，四元数与旋转问题（复数的关键性质）有什么样的关联。在正式写出四元数运算的形式规则之后，哈密顿很快发现，实际上能够用它们来描述三维旋转，方法就是将一个四元数的“虚部”$b\mathrm{i}+c\mathrm{j}+d\mathrm{k}$等同于三维空间中的点$(b, c, d)$。

从四元数到八元数

一旦建立了四元数，下一步就是继续构造其他代数系统。数直线上的实数是一维的，复数是二维的，四元数则是四维的。这个从一到二再到四的推进过程（每次都翻倍）暗示着下一步前行的线索。

年轻的英国数学家，曾就读于剑桥大学的凯莱（Arthur Cayley）也迷上了四元数，受到哈密顿的发现的启示，他阐明了用四元数来定义旋转的方法。他又继续建立了一种八维数的理论，这种数被称为八元数或“凯莱数”。

八元数（ω）具有$\omega = a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k} + e\mathrm{p} + f\mathrm{q} + g\mathrm{r} + h\mathrm{s}$这样的形式，其中的7个虚数（$i, j, k, p, q, r, s$）的平方都是$-1$。八元数中虚数的乘法是根据被称为费诺平面的一种有限几何规则来进行的，费诺平面是由7条线（中心的圆也视为一条线）和7个点（参见第18章）组成的。



用于处理八元数乘法的费诺平面

在这个几何说明中，i、j和k位于中心圆的圆周上，类似于四元数的几何说明，做乘法的方式也是一样的，注意箭头所标注的方向。如果要计算$k \times p$，就要找到包含k和p的线，注意它的方向，结果等于适当方向标记所指的第3个虚数，也就是s。

八元数乘法的顺序也会带来问题，但是八元数系最具有革命性的地方还是在于将3个八元数相乘时，先将哪两个数乘起来结果是有区别的！

越来越多的勇于尝试的数学家接受了这些源于常见运算规则的衍生结果，到19世纪40年代时，人们已经坚信存在着一维、二维、四维和八维的虚数系。根据翻倍的道理，能否更进一步建立十六维的数系呢？很多年后答案才浮出水面：要建立十六维的代数是不可能的。人们最终知道，实数、复数、四元数和八元数系是仅有的可以进行加、减、乘、除运算的系统。

后来人们发现，在航海、机器人和计算机游戏设计等领域（实际上，在很多要用到旋转操作的技术领域）中，都要用四元数来表示旋转。但是对于数学家来说，也许四元数和八元数的真正价值还是在于它们为其他类型的代数开启了更为广阔的前景。

克利福德代数

如果我们不再要求必须有除法，就还可以有别的可能。以19世纪英国数学家克利福德之名来命名的克利福德代数，就是以复数和四元数为基础，但是走上另一条建立八元数的道路。克利福德代数也可以用平方为－1的“虚数”的数目来描述。从本质上讲，它们是为物理学目的而发展起来的，并且在量子理论中找到了用武之地。狄拉克在1928年那篇名为“电子的量子理论”的论文中，在讨论他那著名的电子方程时，就是用虚数来揭示电子自旋的奥秘的。

从虚幻到实用

这段开始于对复数中虚数“i”的不信任的故事，已经明显地改变了数学，现在的数学与虚数出现前那单一维度而且贫乏的样子已

经截然不同了。笛卡儿提出了“虚幻的”这个词，但是由它所带来的复数的美丽构造以及随后产生的各种成果已经改造了数学，对物理学研究产生了重要的影响，而且已经证实这对于各新兴技术领域来说也是不可或缺的。

一旦为数学所接纳，虚数就获得了自主的生命力。它们冲击着代数学的观念和人们已接受的一切。在19世纪，人们发现了它与代数学中其他领域之间的关联，而且这个宝藏依旧熠熠生辉，向我们展示着它的宽广富饶。但是我们必须准备好随时去怀疑那些明显是“合理”的东西，从而实现观念的飞跃。

如果我们指望着将虚数置于包含日常生活所用之数的一维数直线上的话，那么虚数就是不真实的。但是如果我们接受二维表示，那么虚数和复数就和其他任何数一样真实。要是没有它们，今天的数学家就会一筹莫展。而物理学家则从中找到了建立物理理论大厦的最重要的砖石。在这个意义上讲，虚数产生了极其深远的影响。如果柯勒惠支有机会向水晶球中探望一眼，他可能就会看到，正因为我们任由想象力飞翔，才使理性得以发扬。

6 无穷有多大？

——集合论与无穷的变革



深入了解无穷是很困难的。这个概念看起来广阔而深不可测，进而使人感到困惑。怎么会有任何事物具有永无穷尽的数量呢？难道不是所有事物最终都会到达其界限吗？可要真是这样的话，界限之外是什么？很多个世纪以来，无穷困扰着历代哲人，直到最近，当物理学家、天文学家和宇宙学研究者试图理解宇宙时，仍会为之而痴迷。但是数学家为无穷概念带来了新的视界，而且不止于此，他们还借助集合论为无穷提供了新的研究领域。

“无穷”是一个伴随着神秘感的名词。与它有密切关系的一个词是“永恒”即无穷无尽的时间，自古以来我们就对这些难以把握的观念充满好奇。按照传统，出于宗教方面的考虑，包括很多哲人在内，人们认为存在着一个更高级别的力量，掌管着这些不可知的事物。

但数学的需求是不同的，那么，我们如何以数学方式来解释和界定无穷呢？理论上讲，无穷应该是比永恒更容易理解的概念，因为前者少了时间这个因素。

无法获得的无穷与“没有边界的”世界

如果我们问一个小孩子，他所知道的最大的数是多少，小孩子可能会回答诸如100这样的数，或者当他再长大一些时，也会更有自信地回答诸如1万万万亿这样的数。这样的考虑方向是正确的，但是我们总可以回答说，101或者1万万万亿加1就是一个更大的数。片刻之后，他们就可以断定不可能有最大的数，从而一下子跳到“无穷”这个回答以终结所有的疑问。在这里，他们实际上是相当具有革命性的，因为他们把“无穷”当成是一个实际的数。

这可不是公元前3世纪时的古希腊人以及其后众多数学家所持有的无穷概念。他们所考虑的是“潜无穷”。对于亚里士多德和其他一些学者来说，无穷是不可到达的，因此它不是一个实际的数。在表达存在“无穷多个质数”时，几何学家欧几里得说的是“质数的数目比任意给定的数目更多”，以此来引出无限量的概念。

我们使用∞这个特殊符号来表示这种潜无穷，它是由沃利斯在17世纪50年代引入的。一度可能曾使用过ω（读作“欧米茄”），即希腊字母表中的最后一个，不过它的真正来源尚不清楚。为了强调∞不是一个数，数学家们通常不会写$n=\infty$。在表示一个数n变得越来越大时，会用到无穷这个概念，通常他们会说n趋近于无穷，写作$n\to\infty$。

尽管不可到达，无穷在数学家的世界中却始终处于核心地位。伟大的观念来自于对无穷的意义进行公开的探讨，不过这一切都要通过某些变革性的思想才能发生。

让我们从一个可行的无穷定义（也就是我们能在数学教科书中找到的那种定义）开始，无穷“指的是没有界限的量”。这个定义与“越来越大”的想法是一致的，因为整数1, 2, 3, …是可以永远增加而

没有边界的量。任一选定的数总是可以被超过的。但是如果整数有无限多个，那么下面这些小于1的分数又该如何：

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{7}{8}, \frac{8}{9}, \frac{9}{10}, \cdots$$

你确定它们的个数与整数一样多吗？它们可不是“没有边界的”，因为我们知道它们全都小于1。但是这也是对的，因为它们确实还是有无穷多个。

集合论

为了开启无穷之秘，我们实在需要一种比“没有界限”更为精细的无穷的定义，这时该轮到集合论出场了。这方面的数学研究是由康托尔在19世纪70年代开创的。在直觉意义上，集合是事物的总体。例如，正整数作为整体而言是一个集合。为了阐明这一点，我们用大括号{}，并且赋予集合一个记号，对于正整数而言这个记号就是$\mathbb{N}$，记为：

$$\mathbb{N}=\{1, 2, 3, 4, 5, 6, 7, 8, 9, \cdots\}$$

通过这种方法，我们可以把集合看做是单一实体。

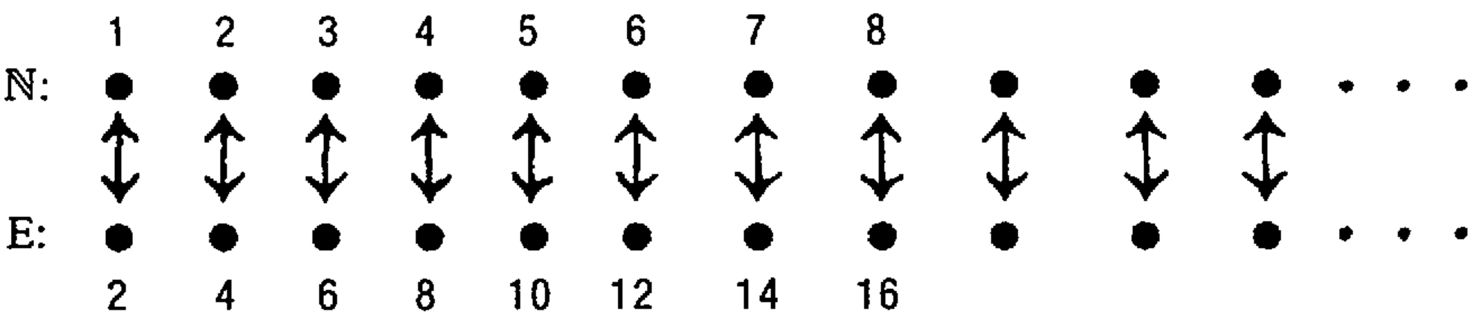
我们还可以讨论$\mathbb{N}$的子集。例如，我们可以有子集$A=\{1, 2, 3, 4, 5\}$，我们称它是一个有限集合，因为它受到其中最大元素（即数5）的限制。这为我们提供了一种不那么恰当的定义无限集合的方法，即不是有限集合的集合。

集合之间的关系

$\mathbb{N}$有一个令人惊奇的性质。让我们问一个天真的问题：偶数，可以表示为子集$E=\{2, 4, 6, 8, 10, 12, \cdots\}$，与正整数一样多吗？直觉

的回答是“不一样多”，因为偶数只占正整数总体的一部分。但是当我们说“一样多”时究竟是指什么，因为无论是对于集合E还是$\mathbb{N}$，我们肯定不能去数其中元素的个数并比较其结果——它们可都是无限集合！

这就是一一对应思想发挥威力的场合了。对于每个正整数，都有一个偶数与之“对应”，反之亦然。这是怎么实现的呢？每个整数对应于它的二倍，比如集合$\mathbb{N}$中的1对应于集合E中的2，2对应于4，依次类推，因而每个数都获得一个对应的数。



集合$\mathbb{N}$与E之间的对应关系

我们还可以确定另一种对应方式，每个偶数都可以通过折半而与一个整数对应，例如集合E中的8对应于集合$\mathbb{N}$中的4。因为两个集合之间存在这种一一对应关系，所以数学家们将它们描述为等势的。

对于正整数的平方，可以表示为集合$G=\{1, 4, 9, 16, 25, 36, \cdots\}$，伽利略（Galileo）进行过同样的演绎推理，因为G和$\mathbb{N}$之间也存在着一一对应关系。某些相当稀疏的集合也与$\mathbb{N}$是等势的，例如集合T，其中的元素是在沿着数直线增加的整数后面依次增加一个零：

$$T=\{1, 20, 300, 4\,000, 50\,000, 600\,000, 7\,000\,000, 80\,000\,000, \cdots\}$$

在$\mathbb{N}$的这个子集中，数与数之间的距离不断增大，但是$\mathbb{N}$与T之间仍然存在着明确的一一对应关系，因而这两个集合还是等势的。

不过这些例子仍旧保持着其矛盾的性质。它们似乎导向某种反直觉的陈述，而公然与常识性的观念（那是古希腊人就已了解的“整

体大于部分”）相抗衡。

在希尔伯特旅店登记

“希尔伯特旅店”，因其提出者——数学家希尔伯特而得名，它是一个思想实验，能够揭示出无限集合的这种不同寻常的性质。这所想象中的旅店拥有无穷多个房间。我们假设到某个傍晚7点时，旅店中“住满了人”，也就是说，对于每个数n，n号房间都住着人。到晚上8点时又来了一位新客人。希尔伯特该怎么做？他应该告诉这位客人“客房已满”还是为他另找个住处呢？他变回了数学家，为这个难题找到了一种方法，开始调度已有的房客。他请住在1号房间中的房客搬到2号房间中，2号房间中的房客搬到3号房间中，以此类推，每位房客都搬到隔壁的下一个房间中。新来的旅客住到腾出来的1号房间中。问题解决了！

晚上10点，更富戏剧性的事情发生了。火车带来了无穷多个想要入住的旅客。希尔伯特镇静如常。他再次调度起房客来，这次是将他们的房间号翻倍。于是，住在1号房间中的房客搬到2号房间中，住在2号房间中的房客搬到4号房间中，其余同理。经过这番调度，所有奇数号的房间腾了出来，希尔伯特愉快地将它们分配给新来到的旅客。即便此后再有无穷多个客人来到，他仍然可以通过重复同样的操作而安排他们入住。希尔伯特旅店的接纳能力是无限的。

提出者希尔伯特是一个做事很有条理的人，他注意到给所有留住在希尔伯特旅店的客人编号是有用的。因此，他参照房间号写下入住者的名字。这确实会是一个无穷序列，但是他至少能够说明每个房间的入住者是谁。

编号排序，然后数两遍

希尔伯特提出的这种构造编号序列的思想，已经被证实在数学家对于无穷的理解中占据中心地位。集合 ℕ 可以看做是一个序列1, 2, 3, 4, … ℕ 的子集可以置入序列中，但是，是否还有更大的集合能够编成序列呢？答案是“是”——比如包含零负整数和正整数的整数集合 ℤ：

$$\mathbb{Z} = \{\cdots -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, \cdots\}$$

怎样才能将 ℤ 编成一个序列呢？ℤ 沿着数直线从 $-\infty$ 向 ∞ 延伸，怎么可能将它们编成序列呢？一种方法是选择0作为起始数，1作为第二个，-1作为第三个，以此类推，将增加和减少的数交替排列，得到的序列是：

$$0, 1, -1, 2, -2, 3, -3, 4, -4, 5, -5, 6, \cdots$$

更不可思议的是，我们甚至能将所有分数（用Q来表示）编成一个序列。这个序列不是按照数值大小来排序的，但却肯定是一个序列。在尝试过程中，我们可能会先试着写出所有整数（写时带着分母1)，接着写出分母为2的分数，再写分母为3的分数，以此类推：

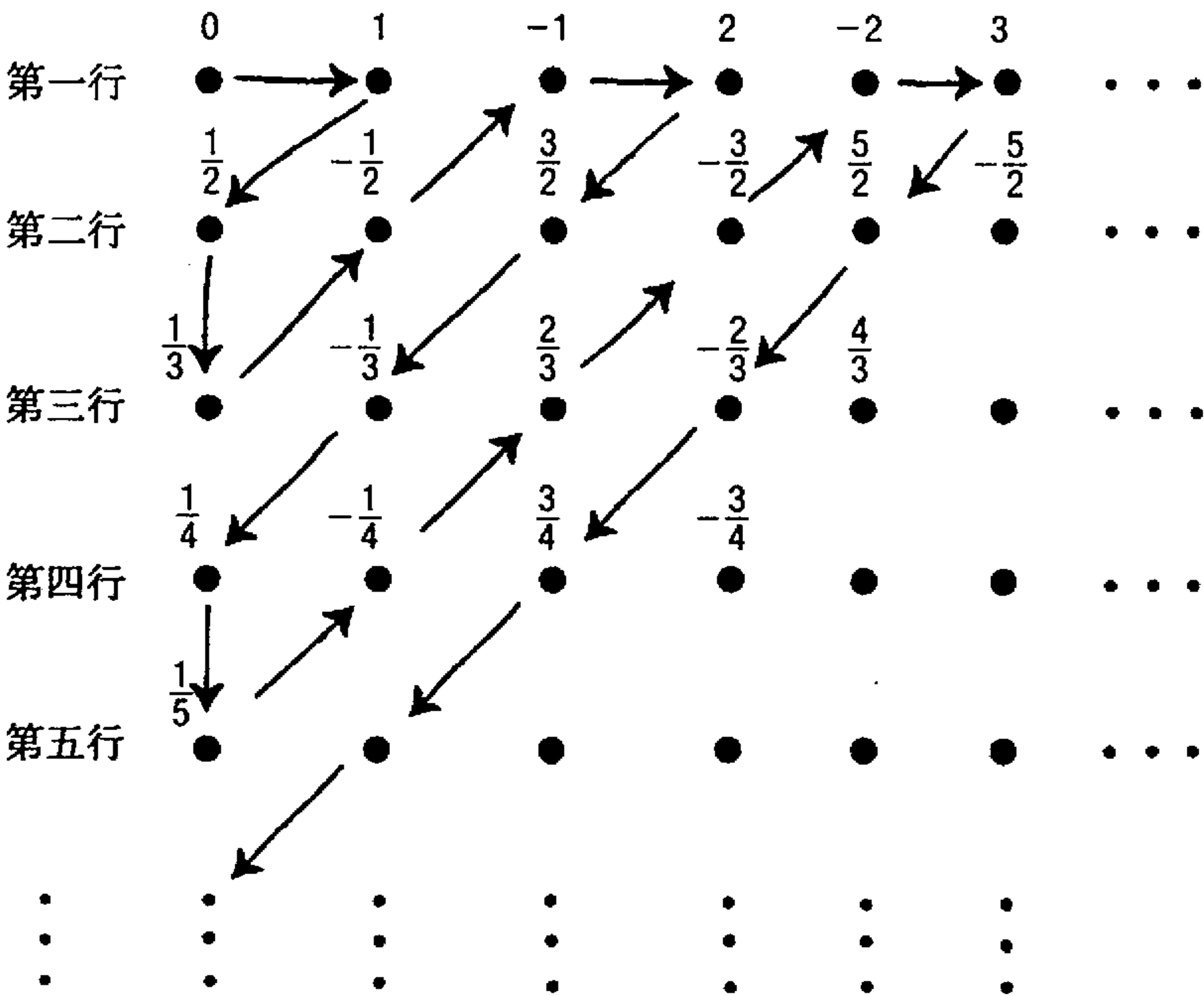
$$0, 1, -1, 2, -2, 3, -3, 4, -4, 5 \ \cdots$$
$$\frac{1}{2}, -\frac{1}{2}, \frac{3}{2}, -\frac{3}{2}, \frac{5}{2}, -\frac{5}{2}, \frac{7}{2}, \cdots \quad \frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}, \cdots$$

这是一个编号序列吗？我们能够分清序列中的第一个、第二个、第三个甚至第四个是谁，但是$\frac{1}{2}$处于什么位置？麻烦在于，我们没法跳过不同分类之间的省略号“…”。不过也不是一无所获。我们可以先将Q中的分数排成一个二维的阵列。阵列的第一行是整数，下一行是分母为2的分数（忽略像$\frac{4}{2}$那些重复出现的数，$\frac{4}{2}$等于2，已

经在上一行出现过了）。在第三行中，我们置入所有分母为3的分数，仍像刚才那样忽略重复的数。

现在我们可以构造一个序列了，从左上角的0开始，移动到1，但接下来沿对角线方向移动到$\frac{1}{2}$（如图）。跳动着“对角线舞步”，我们就能写出如下序列：

$$0,\ 1,\ \frac{1}{2},\ \frac{1}{3},\ -\frac{1}{2},\ -1,\ 2,\ \frac{3}{2},\ -\frac{1}{3},\ \frac{1}{4},\ \frac{1}{5},\ -\frac{1}{4},\ \frac{2}{3},\ -\frac{3}{2},\ -2,\ 3,\ \frac{5}{2},\ -\frac{2}{3},\ \frac{3}{4},\ \ldots$$



排成二维阵列的分数集合Q

每个分数都处于序列中的某个位置，而且我们总可以分清它在哪里，进而确定它所对应的整数。这个序列不是按照数值大小排成的，但它确实是一个序列。

ℕ和Q都是无限集合，而且它们都能写成序列，因此二者是等势的。此外，所有能够排成一个序列的数集都是与ℕ等势的。

不可列的集合

关于数学家们所称的实数ℝ，那些可以沿着数直线排列的数所组成的集合，我们能说些什么呢？ℝ中包括常数π和其他的无理数（参见第4章），它被称为实数*连续统*，因为它能够无空隙地填满数直线。与集合Q一样，连续统ℝ也是无限集合，但是我们能够将ℝ中所有数排成一个序列吗？如果可以的话，那么ℝ就和Q等势。

19世纪70年代时，康托尔发现了数学中最令人惊讶的结果之一。事实上，他说明实数*不可能*写成一个序列。他的论证是这样的。每个实数都有一个十进小数表示，例如，我们可以有一个数34.896 749 4…如果想要构造一个与它不同的数，我们可以不改整数34，而将小数点后第一位数字改成1，从而得到34.196 749 4…这个数与原数34.896 749 4…只有一位小数不同，但足以保证它是不同的数。（如果小数点后第一位数字本来就是1，我们只需将它换成2即可。）

现在假定已经存在一个由实数ℝ排成的完整序列。利用已经描述过的方法，我们能够构造出这样一个数：它与序列中第一个数的第一位数字不同，与序列中第二个数的第二位数字不同，以此类推，结果是我们构造出来的这个数不可能出现在序列中的任何位置，因为它与其中所有数都是不同的。

无可辩驳的结论是，不可能存在一个由实数ℝ排成的完整序列。

通过将分数添加到整数集合ℕ中，我们得到了更大的集合Q，但是它与ℕ具有同种类型的无穷属性。而当我们继续将无理数添加到分数中时，则会得到一个具有*更高阶无穷*属性的集合ℝ。这是革命性的。无穷不只有一种而是有两种，Q之无穷的阶与ℝ之无穷的

更高的阶。康托尔的理论与一些久负盛名的数学家的观念相冲突。

无穷的阶与连续统假设

对于无穷的不同阶，康托尔引入基数的概念作为区分它们的方法。大体上讲，一个集合的基数就是它的大小，比如集合$\{a, b, c, d, e\}$的基数就是5，因为其中有5个元素。但是对于无限集合我们该怎么做呢？康托尔将$\mathbb{N}$的基数记为$\aleph_0$，其中$\aleph$（读作“阿列夫”）是希伯来字母表中的第一个字母。$\aleph_0$这个符号，读作“阿列夫零”，自从康托尔时代以来就一直没有改变。集合$\mathbb{R}$的基数记作c，来自于“连续统”，由于$\mathbb{N}$的无穷比$\mathbb{R}$的无穷小，因此我们可以写成$\aleph_0 < c$。

> “对无穷的畏惧是缺乏远见的表现，它会破坏我们看到实无穷的可能性，即使实无穷的最高形态是使我们得以创生和维系的根本，而它的次级超穷形态也是无所不在甚至深植于我们的头脑之中。”
>
> **格奥尔格·康托尔，《论实无穷》**

我们能否在$\aleph_0$和c之间再插入一个基数呢？在常见的算术问题中，有类似$\frac{1}{2} < \frac{4}{7}$这样的不等式，那么在任意两个分数之间总可以插入另一个分数。要实现这一点，一个简单的方法是将分子和分母分别加起来，得到$\frac{1+4}{2+7} = \frac{5}{9}$，从而发现有$\frac{1}{2} < \frac{5}{9} < \frac{4}{7}$。但是如果我们只考虑整数，那么在诸如2＜3这样的不等式中，是不可能再插入另一个整数的。

连续统假设是这样一个假设：在$\aleph_0$和c之间没有任何基数，或者换一种方式来讲，$\aleph_0$之后的下一个基数就是c。康托尔试图证明连续统假设，但是历经多年他仍然无法做到。20世纪的数学家哥德

尔相信这个假设是错误的，但是也无法证明这一点，不过后来，他和他的美国同事科恩证明，在标准的集合论框架内，连续统假设既不能被证实也不能被证伪。这相当于说，连续统假设独立于通常用来刻画集合论的公理集。

这是一个重要的转折点。我们已经知道，存在着很多种不同类型的几何学，而现在，在集合论领域中，连续统假设的独立性使我们获得了拥有不同类型集合论的可能性。

大批的无穷

关于无穷的本质，康托尔还有着其他一些发现，最著名的当属他创造的*超穷*基数。从$\aleph_0$出发，他构造出一个具有更高阶无穷的集合。以$\mathbb{N}$为起点，他构造了由$\mathbb{N}$的所有子集组成的集合$\mathbb{N}_1$，并且证明它具有比$\mathbb{N}$更高阶的无穷。将$\mathbb{N}_1$的基数记为$\aleph_1$，从而有$\aleph_0 < \aleph_1$。对$\mathbb{N}_1$重复同样的构造，产生具有基数$\aleph_2$的集合$\mathbb{N}_2$，这个过程可以不断重复，产生一个完整的“阿列夫”序列，它们具有越来越高阶的无穷：

$$\aleph_0 < \aleph_1 < \aleph_2 < \cdots < \aleph_n < \cdots$$

通过说明怎样构造出每一项都具有不同阶无穷的阿列夫序列，康托尔向我们解释了构造出一个由无穷组成的无穷序列的方法。现在可以建立一种算术理论，在其中有可能用基数进行加法和乘法，从而为逻辑学家和数学家们提供一片广阔的用武之地。

我们的话题开始于对无穷之本质与它究竟有多“大”的好奇。“无穷”这个词的含义，不仅有须弥之广，也可以有芥子之微，当我们描述某种事物无限微小时便会概括出这个词。在数学知识中，它反映为整数可以不断地延伸，而在另一个尺度上，我们可以没完没了地创造出与数1差距越来越小的分数。

而在深入探索无穷的过程中，我们又发现存在着远比整数更多的数集。不仅如此，我们还发现，任何无穷的概念之上都可以有一个更高阶的无穷。我们实现了这样一种可能性，无穷的概念不止一个，而是有无穷多个。

7 平行线在哪里相交？
——新几何学的诞生

于“平行线在哪里相交”这个问题，简单的答案就是它们绝对不会相交，因为平行线的定义就是不会彼此相交的直线。平行线是历史上出现过的最多产的数学定义之一，它可以追溯到公元前300年时亚历山大里亚的欧几里得。但是故事并未在那时就结束。为给平行线性质提供一个证明，所做的种种尝试为几何学开拓出一片新的天地。它们不仅改变了我们对于数学本身的认识，还为我们研究空间的物理学提供了方法。

两千多年以来，直到19世纪末时，欧几里得的《几何原本》几乎就是几何学的同义词。要是你想研究几何学，那就先要研究欧几里得的论述，当你沿着《几何原本》所指明的道路前行时，你才能保证自己是在跟随着数学泰斗在研究几何学。几何学只有一种，就是欧几里得在他那本“神圣著作”中所记录的那一种。如同对待《圣经》一样，他的信徒们可以引述其中的章节和词句。

不仅如此，由于《几何原本》的13卷内容是按照严格的逻辑来设置的，因此它也成为应该怎样做数学的典范。你先定义自己要用的名词，提出自己的假设，接着利用证明出来的定理系统地建立一整套知识。欧几里得用到了其他数学家的成果，但是他的天才在于给出了这种合乎逻辑、条理分明的组织方式，例如，关于直角三角形的毕达哥拉斯定理在《几何原本》中是第1卷的第47个命题，或记

为1，47。

平行公设的奥秘

《几何原本》中的每条命题都是基于5条“公设”而得到严格证明的，这些公设是欧几里得构造出来的假设，在他看来，它们都是无可争议的论断。尤为特别的是，公设中谈到了点、线和角的基本性质，以及它们相互作用的方式。

前4条公设如下。

(1) 从任意一点到另一点可以做直线。

(2) 一条有限长的直线段可以不断延伸，成为一条直线。

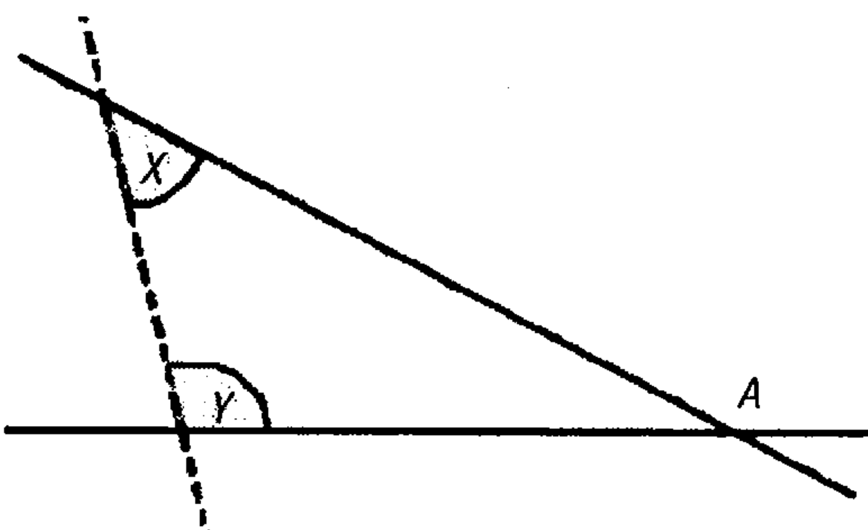
(3) 以任意圆心和半径可以做一个圆。

(4) 所有的直角彼此相等。

到目前为止，都是些直截了当的陈述。但是对于欧几里得第五公设，却存在着令人不安的疑惑，这条“平行公设”看起来相当地冗赘而曲折。

(5) 一条直线与两条直线相交，若同旁内角之和小于两倍的直角，则当两条直线无限延伸时，会在同旁内角之和小于两倍直角的那一侧相交。

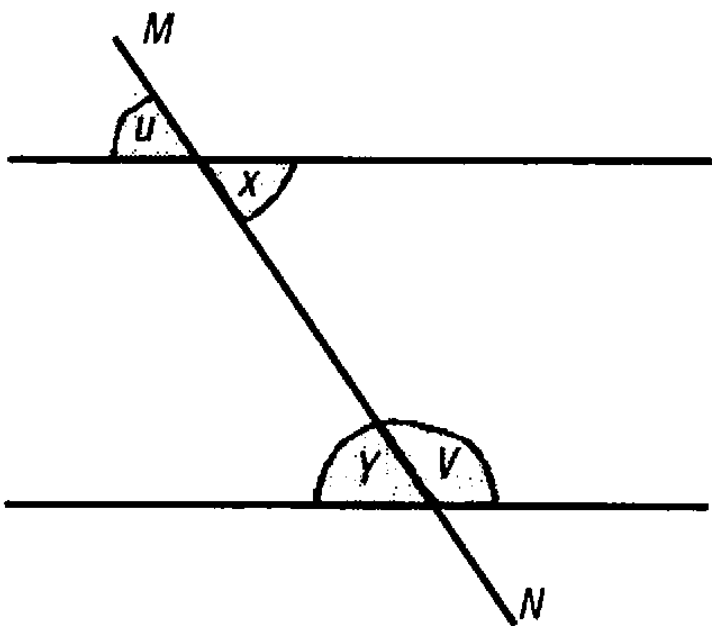
欧几里得需要一种能够把握平行线的方法，这就是构成第五公设的出发点。但是它能够用其他4条公设证明出来吗？如果可以的话，就应该将它从我们必须提出的假设中剔除出去。很多人进行过尝试却没有收获，但这些探索最终将



欧几里得的“平行公设”

改变几何学，进而改造我们对于物理世界的理解。

尽管前4个公设简短而直接，而第5个看来却更像是一篇短文，而且欧几里得不像是在说什么是平行线，而更像是在说什么不是平行线。用图示的方法也许可以更好地阐明这一点，图中，一条虚线与两条实线相交，形成夹角X和Y，它们的和＜180°（即两倍的直角），因此两条实线最终一定会在与两个夹角处于同侧的某点(A)处相交。



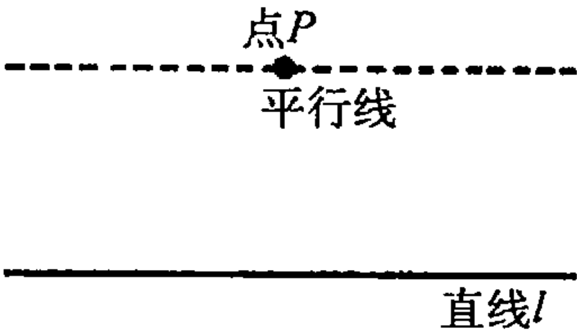
欧几里得的命题29

我们可能会奇怪，欧几里得为什么要在他的逻辑框架中引入这样一条笨拙的公设，但这种做法恰好表明了他的天才。在他的几何论证中，第五公设是必不可少的，但他只在自己被逼入角落的情况下才迫不得已地使用它。在不使用它的情况下完成了尽可能多的证明之后，第五公设在《几何原本》中的首次亮相要等到命题29的证明。

> 一条直线[MN]与两条平行线相交，形成彼此相等的一对内错角[$X=Y$]，靠外部的角与内部的对顶角相等[$U=X$]，因而同旁内角之和等于两倍直角[$X+V=180°$]。

很多个世纪以后，在爱丁堡工作的数学家普赖菲尔于1795年出版了《几何原本》的一个版本，对平行公设给出了一个等价但更为直觉性的表述。

> 过一个给定的点刚好能作一条直线平行于给定直线。



普赖菲尔的公设

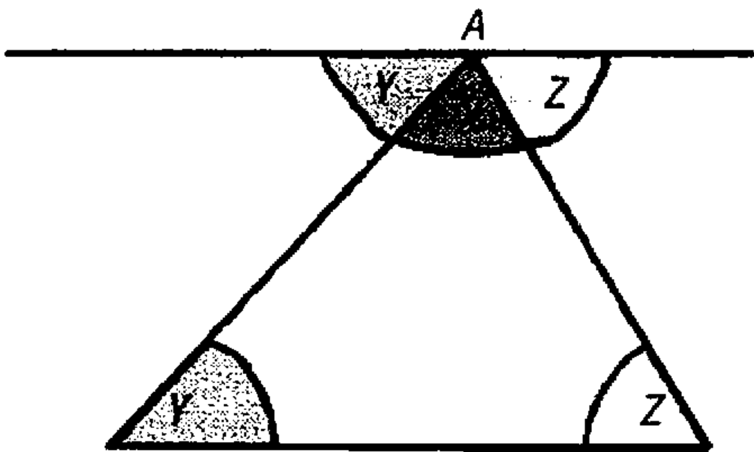
普赖菲尔公设的简明性使其得到普及，此后各版《几何原本》的编纂者把它看做是缓和欧几里得那严格到苛刻之风格的一种方法。但是，无论是用欧几里得的还是普赖菲尔的语

言，这条公设始终缺乏一个令人满意的证明。

寻找一个证明

多年以来，证明平行公设这个问题难倒了很多大数学家。潜在的易犯错误之一是落入循环论证的陷阱，也就是说，试图用实际上就是该公设的前提去证明它，换言之，等于假定该公设已经得证。

欧几里得已经证明三角形的内角和是180°，不过毕达哥拉斯所给出的证明更好懂些。对于任一给定的三角形，直接应用普赖菲尔公设，我们可以做出经过“顶点”（与底边相对的点）A的唯一一条平行线。利用欧几里得的命题29中关于内错角的结果，我们在图中标记角Y和Z以及剩下的最后一个角X。现在我们可以用两种方式来看待这些角。从一个角度来看，位于同一顶点附近的这些角构成了一条直线，因此$X+Y+Z=180°$；从另一个角度来看，$X+Y+Z$就是三角形的内角和。证毕。



三角形的内角和＝一个点周围的各角之和

*逆陈述*也是可以证明的：如果三角形的内角和是180°，并且过顶点A所作的直线与三角形的两边分别形成角Y和Z，那么这条直线与三角形的底边平行。

当一个结果和它的逆都得到证实后，就称它们是*等价的陈述*。重点在于，欧几里得的平行公设与三角形内角和是180°这一陈述是等价的。任何人想要证明第五公设都必须要了解它的各种等价形式，因为如果在证明中使用了其中任何一条，那么其论述就是循环论证，因而是无效的。从第五公设各个等价陈述的伪装形式中，可以看到

这一公设的精微意蕴。它们看起来是如此隐蔽，甚至都没有提到“平行”这个词。

在任意直角三角形中，毕达哥拉斯定理成立。

存在着一个每个内角都是直角的四边形。

证明平行公设折磨了勒让德三十多年。作为一个一流的学者，他编写的教材《几何学原理》（*Eléments de géométrie* 1794）直到一个世纪之后还在使用。但是他也掉进了循环论证的陷阱，因为他预设了三角形内角和是180°。

欧几里得“获得平反”

在所有解决这一难题的尝试中，18世纪数学家萨凯里的贡献是格外重要的。他指出，在研究经过一给定点的直线与一给定直线的关系时，从逻辑上讲只有3种情况需要考虑。

(1) 经过给定点有且仅有一条直线平行于给定直线（这就是欧几里得的几何学）。

(2) 经过给定点有多于一条直线平行于给定直线（也就是说，平行线不唯一）。

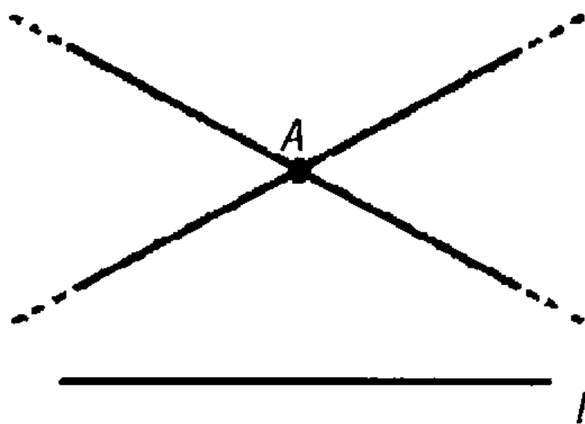
(3) 经过给定点不存在平行于给定直线的直线（也就是说，平行线不存在）。

用第二条假设替换欧几里得的平行公设，萨凯里得到了一些奇怪的定理，他无法解释更从未见到过这些奇怪的定理。用第三条假设替换后，他得到了一些自相矛盾的结果。对他而言，摆脱这种异常境况的唯一方法就是接受第一条假设，即平行公设。他满怀自信地将这一论断发表在自己的著作《欧几里得获得平反》（1733）之中。但是他并没有提出一个无可辩驳的证明。

从一条平行线到很多条平行线

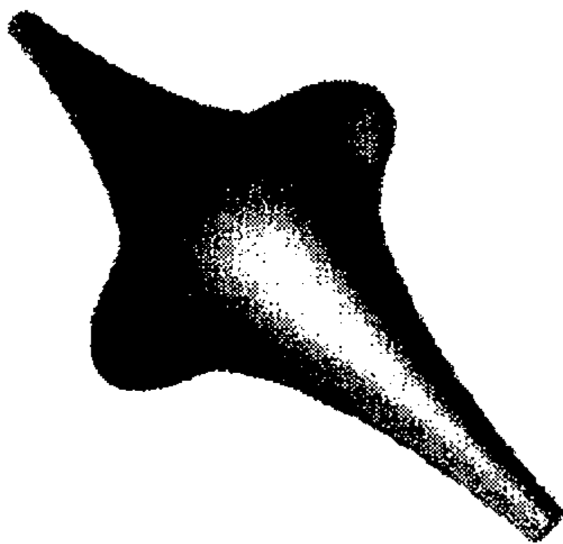
到了19世纪30年代，年轻的俄罗斯数学家罗巴切夫斯基得到了一个结论：既然那些最伟大的数学家们都没能证明平行公设，为什么不用其他假设来替换平行公设，而得到其他的几何学呢？

罗巴切夫斯基选择了萨凯里的第二条公设，即假设平行线不唯一。这意味着，若给定一条直线（l）和不在该直线上的一个点（A），存在着多于一条经过该点且不会与给定直线相交的直线。从常识的角度来讲，这似乎是不可能的。在一张平滑的纸上，过一个给定点，任意画多于一条直线，肯定无法得到多于一条直线平行于已有的给定直线。在经过一定的延伸后，任意其他直线都一定会与给定直线相交。不仅如此，匈牙利的雅诺斯·鲍耶也于同时发表了与欧几里得处理方法不同的结果，他也认同罗巴切夫斯基的结论。



罗巴切夫斯基的平行线

这一想法的关键在于所关注的不只是平面而是曲面。那不是一个具有正的常数圆度或曲率的球面，而是一个具有负的常曲率的曲面。实际上，罗巴切夫斯基和鲍耶创造了“伪球面”上的几何学，或者用更广为人知的名称来说——双曲几何。伪球面上的直线是“测地线”，也就是两点之间的长度最短的曲线。这个图形归功于贝尔特拉米，它为一种非欧几何学提供了一种现实模型。


伪球面与其上的多条平行线
（罗巴切夫斯基的几何学）

新型几何学中的定理与欧氏几何学中的全部定理都没有矛盾，因为罗巴切夫斯基和鲍耶保留了欧几里得的其余4条公设。证明时不需援引平行公设的那些定理（截至命题29之前的那些）还可以保留

下来。但是有一点是会有影响的，即三角形内角和定理。在欧氏几何学中，如同我们已经看到的那样，三角形内角和是180°这一事实与平行公设是等价的。而罗巴切夫斯基和鲍耶用新的公设取代了它，三角形的内角和就不再等于180°。

事实上，令人吃惊的是，在新型的几何学中可以严格证明，一个嵌入伪球面的三角形的内角和小于180°。不仅如此，一个三角形的内角和依赖于它的面积，而且，三角形的面积越大，内角和就越小，反之亦然。这种关系与欧氏几何学中的完全不同，在后者中一个三角形的面积与它的内角和没有关系。欧氏几何学中任意大小的三角形内角和都是180°。

罗巴切夫斯基是喀山大学的一位教授，而鲍耶是一位军官，他们在19世纪30年代初发表了各自的成果。但他们的成就的重要性没有得到应有的认可，而是被忽略了。这时，伟大的卡尔·弗里德里赫·高斯出场了——尽管他的著作是在死后才发表的。高斯在这个领域的研究并不为人所熟知，他已经得到了类似的成果，但却畏惧那些盲从于欧几里得之人的批评。因此他没有发表自己的成果。直到1855年他去世之后，关于非欧几何的私人信件发表出来，其他数学家才因他的威望而确信应该同时去读罗巴切夫斯基与鲍耶的相关文章。

从一条平行线到没有平行线

欧几里得的全知全能地位遭受了打击，但是更糟的还在后面。19世纪晚期，伯纳德·黎曼给出了比罗巴切夫斯基和鲍耶更为激进的改进方案。他改动了欧几里得的第2条公设“一条有限长的直线段可以不断延伸，成为一条直线”，将它变为“所有直线都只有有限长度但没有末端”。尽管这一陈述看起来是那样地神秘，在某些几何学

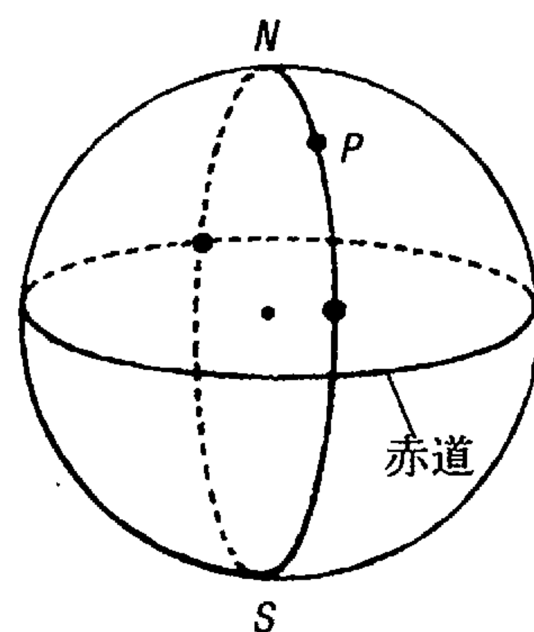
中被看做是一条“直线”的一个圆就符合这个要求。它的长度（也就是它的周长）无疑是有限的，当然一个圆不会像常见的线段那样有末端。

黎曼所做的还不止于此。他还改动了平行公设，实际上他选择的就是萨凯里的第3个方案，即“经过给定点不存在平行于给定直线的直线”。黎曼是怎样使平行线消失的呢？相比于罗巴切夫斯基和鲍耶以伪球面作为承载其新型几何学的模型，黎曼考察的是真实球面。

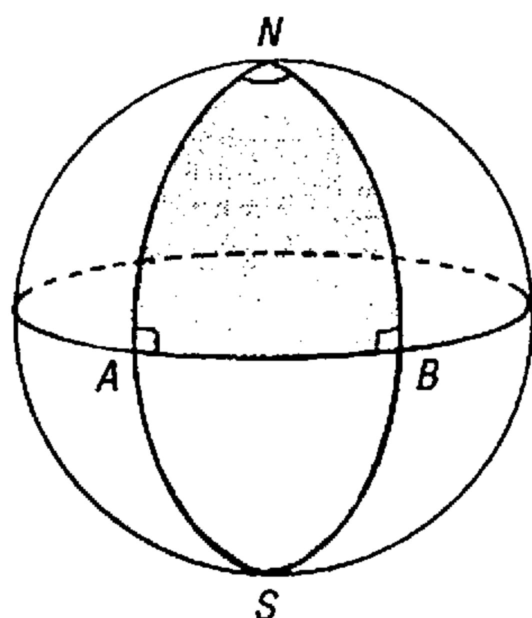
球面几何学本身并不是全新的，它曾为古希腊人所研究，而且一直以来就对航海活动有着重要的意义，因为我们的世界就是球形的。在黎曼几何中，球面上的“大圆”（即圆心位于球心处的圆）就相当于直线。经度线是这种几何学中的直线，但纬度线不是。

为了理解这种几何学中如何解释平行公设，我们可以将赤道（一个“大圆”）看做是给定的直线，并且想象地球上不在赤道上的某处的一个点（比如说巴黎）作为不在直线上的点P。过P点的所有直线（“大圆”）都与赤道交于两点，因此在这种几何学中，我们在球面上无法找到任何经过P点又不与给定直线相交的直线。在这种几何学中不存在平行线。

欧几里得在证明时没有使用公设2或公设5而得到的那些定理，对于黎曼的球面几何学依然有效。但相对于欧几里得正统观念而言，还是产生了更多惊人结果。如果考察球面上以北极N为顶点、以赤道为底边所形成的任意三角形，就会发现三角形NAB的两个底角都是直角。由于北极顶点处还有一个角，因此在黎曼几何中，三角形



球面上没有平行线



在球面上，三角形内角和超过180°

内角和超过180°。

与在罗巴切夫斯基和鲍耶的几何学中一样，内角和依赖于三角形的面积。但在这里，具有较大面积的三角形拥有较大的内角和。比如，当图中位于北极顶点的角增大时，三角形的内角和随之增加，很明显，三角形的面积也会明显跟着增加。

> 牛仔们有一种捆绑公牛或不驯服的野马的方法，将其牢牢固定，使它们既不能移动也不能思考。这是一种极大的束缚，而欧几里得对几何学所做的也是如此。
>
> E.T.贝尔
> 《对真理的探寻》（1934）

借助更为广阔的历史视角，我们实际上可以看到，由黎曼所推动的非欧几何学在人类面前闪耀已有很多个世纪。公元100年左右，亚历山大的梅内劳斯实际上已经在这种“椭圆”几何中开始进行计算了。但是人类有时候会受限于平地生活，尽管我们知道世界是圆形的，它只是在局部看起来是平坦的，并且通常使我们产生“平坦”的体验。

几何学的清单

新的几何学并未导致任何矛盾的这一事实意味着平行公设是独立于其他公设的。绝没有人能够证明它，之前提出的所有对它的证明都必然是错误的。数学家们已经证明，他们能够自由地选择别的公设来替换平行公设，结果是推演出不同类型的几何学。

欧氏几何学与新型的非欧几何学可以分别进行研究，但是德国数学家克莱因找到一种有用的好方法，将它们加以整理，以便阐明

其间的关联。

	几何学的类型	平行线	三角形内角和
罗巴切夫斯基，鲍耶	双曲几何	存在任意多条平行线	小于180°
欧几里得	欧氏几何	只存在一条平行线	等于180°
黎曼	椭圆几何	不存在平行线	大于180°

克莱因对几何学的分类

不过，现在数学家们需要的是一种研究空间的方法，它在局部与欧氏几何一致，而在更大的尺度上则更为复杂。1854年，黎曼进行了名为“几何学的基础假设”的演讲，在演讲中他引入了“流形”的概念，这是一种复杂的空间，但是用比喻的方式来讲，在微观上看起来是平坦的（参见第17章）。

宇宙的几何学

这种新的“黎曼几何”的出现，对于分析时间与空间相互作用的方式具有重要的意义。超大质量的存在会以何种方式改变其周围的几何学，这就需要“流形”的概念，因为长度、角度和曲率的度量都会随着位置而变化，而不是像球面那样所处的曲率都相等。不出意料，爱因斯坦十分推崇黎曼的思想，并称如果不是了解到这些内容“我就绝对不可能建立相对论”。

在爱因斯坦之前的两千多年里，欧几里得的《几何原本》一直是极其出色并且也历久不衰的优秀成就。他的平行公设在经历了充分而强有力的挑战之后，最终导致了新型几何学的诞生。但是新几何学并没有导致欧氏几何“出错”，尽管从现代视角来看其中的逻辑过程使用了未经宣称的假设。从实用的角度来看，我们还能够继续利用欧氏几何学知识来架设桥梁和进行建筑投资。我们在日常生活

中对于平行线的理解仍然是“欧几里得式的”，而且当我们处理平面问题时，还是会满怀信心地求助于那位伟大的几何学家。平行公设的故事导致了新型几何学的产生。当人类试图恰当地理解几何学以及发现物理空间的真正本质时，黎曼几何和“流形”的概念就是必不可少的。

8 什么是宇宙的数学？——微积分奇迹

18世纪的作家波普在向一位伟大同胞的致辞中写道：“让牛顿出现！于是世界一片光明。”也许他还应该在这番颂词之中加上德国人高特弗雷德·莱布尼兹，因为是这两位自然哲学家和数学家一起开创了我们简称为微积分的计算方法，为天文学计算以及其他很多领域提供了精确的研究手段。历经多年琢磨和砥砺，已经证明这门学问对于整个科学领域来说是无价之宝。它的应用还在不断拓展，涵盖了包括人口学、经济学、医学、统计学和量子物理学等诸多学科。

微积分的发现，缘起于人类仰观天象的活动，以及试图理解太阳系乃至其外部之运行机制的渴望。在牛顿和莱布尼兹的手中，它被成功地应用于计算和对行星运动的描述。自他们的时代以来，它逐渐获得了严格的基础，并且在人类渴望去描述、度量和理解变化的任何场合找到了用武之地。

作为数学家的牛顿同时也是科学家、实验研究者和天文学家。当他还是一个年轻人时，就深深地迷上了行星运动和那些能用来观测天象的科学仪器。才只有30岁时，他就因设计出一种反射望远镜而在伦敦引起了一场轰动，正是由于这一成就，他被推选进入英国皇家学会。

从“向心”力（物体所受到的来自一个点的吸引力）的性质出发，牛顿建立了伟大的万有引力理论和著名的运动三定律。以此为基础，他用微积分工具推导出行星绕着太阳以椭圆轨迹运动的结果。开普勒曾经基于实验数据而提出了行星运动椭圆轨迹的理论。牛顿则到达了更高的层面，利用数学的“变化”理论将其推导出来。在微积分方法中，牛顿找到了开启宇宙奥秘的一把数学钥匙。

流动的世界

微积分（Calculus，按照惯例，这个词的首字母一般要大写）建立在如下观念的基础之上：我们身边的世界是流动的、不固定的，在不断发展中的。每一分钟都与前一分钟不同。整个世界都处于流动的状态中。公元前5世纪时，哲学家赫拉克利特总结出这样一句话：“一切都在流动，没有任何事物保持静止。”再补充一句：“你无法步入同一条河流两次。”

我们熟知这一点。从被抛向空中的小球所处的高度，到我们发射的火箭到地球的距离、人口数量的多少、流行病的传播，再到受磁体存在影响的电流大小——几乎一切现象都要涉及空间和时间中的变化。微积分的重要思想是我们观测到的任何变化都依赖于某些变量的值。换言之，被抛向空中的小球所处的高度在很大程度上取决于它离手后所经历的时间。（当然，小球的高度可能还取决于其他一些因素，比如空气阻力，不过出于实用考虑，某些因素是可以忽略的。）

在研究变化时，最明显的变量是时间，但是，微积分能够在任何情况下处理所有类型的变量。最早能够被看做是微积分问题来考虑的问题，可以追溯到古希腊人，但亚里士多德和赫拉克利特是以哲学方式而不是以数学方式来考察变化的。伽利略和某些中世纪数

学家确实是沿着数学道路在探索，但这一主题真正的现代开端还得归功于两大人杰——牛顿和莱布尼兹。他们都享有“首先”发现微积分的荣誉，而实际上，两人是在彼此独立的情况下用不同的数学符号形式取得各自的成果的。牛顿的符号在涉及运动的问题中更有帮助些，但是事实证明，它没有莱布尼兹的符号那样普及，后者在今天更受人青睐。

由这两人所发现的微积分，包含着互补的两个方面：“微分学”和“积分学”。实际上，它们构成同一个硬币的两面，具有互逆关系。前者考虑的是“拆开”，后者则是“合并”。

量度变化

微分学的核心目标是度量变化率，即变化过程发生得多快或者多慢，我们将它称为“导数”。寻找导数的过程称为“微分”。

此外，微分学还关心特定时刻的变化率，也就是说，瞬时变化率。因此，它与平均变化是大为不同的。例如，假设从城镇A到城镇B的距离是330英里，一辆小汽车用11小时走完这段路，则它的平均速度是330除以11，结果是30英里每小时。但是这并没告诉我们在旅途中任一时刻的速度。在某个时刻，小汽车可能是静止的，比如等交通灯时，而在另一时刻，它也可能以每小时70英里的速度疾驰在高速公路上。

用代数方法来看，我们所考察的量可以表示为y，在将小球抛向空中的例子中，它就是小球距离地面的高度。研究y变化的速率，是微分学中的核心问题。

牛顿曾专门研究瞬时速率，即距离在特定时刻的变化率。他还考虑过扩展的或称高阶的变化率（瞬时加速度）的概念，即速度的变化率。牛顿使用流数方法来重点考察瞬时变化。用他的语言来讲，

若y最终流动了，则称为一个“变数”。这个变数改变的速率，他称之为“y的流数”（用他的符号记为$\dot{y}$）。再向前一步，$\dot{y}$的变化率记为$\ddot{y}$。于是，当y表示距离时，$\dot{y}$就是速度（或速率），而$\ddot{y}$则是加速度。

不过，还是莱布尼兹的微积分符号能够帮助我们更好地理解瞬时速度。在将小球抛向空中的例子中，我们最终必须要想象两个点：一个特定时刻（x），也就是我们希望了解其速度的那个时刻；以及在随后延长时段中的某个时刻。两个时刻之间的时间段用Δx来表示（这是一个整体符号而不是$\Delta \times x$，读作“德尔塔x”）。在这个延长的时间段中，小球将要经过的距离记作Δy（读作“德尔塔y”）。

这段延长的时间中的平均速率（用走过的距离除以经过的时间）用$\frac{\Delta y}{\Delta x}$来表示。用$\Delta x$表示的时间段越小，$\frac{\Delta y}{\Delta x}$就会越接近于$x$时刻的瞬时速率。

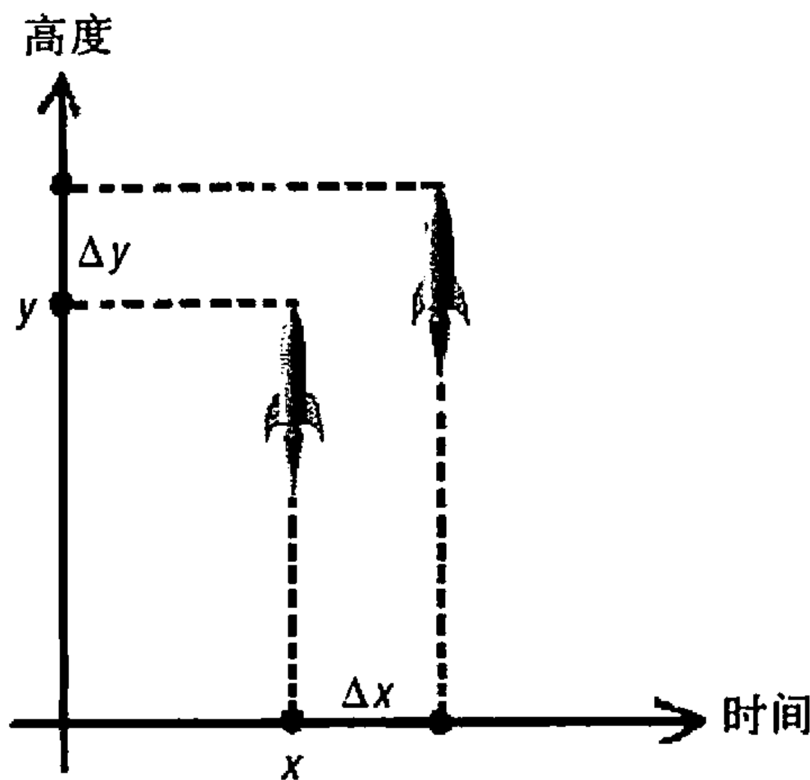
对加速度的考察

为了更好地欣赏微分学中各种技术细节如何构成一个有机整体，让我们来考虑火箭发射的例子。这与将小球抛向空中的例子相似，但我们还要努力理解一下火箭推进的特性。

现在，我们要研究的是火箭在x分钟后的瞬时速率。如果这个速率是常数（比如说每分钟10英里），那么它就是在匀速运动，这个速率也就是任意时刻的瞬时速率。但是火箭的运动并不是那样的。它的起始速度是零，最初运动得很慢，接着在加速过程中速度逐渐增大。

实际上，火箭行走距离与历经时间之间的关系要依赖于与该火箭有关的各种因素，例如火箭上装载燃料的数量、空气阻力以及地

球引力随火箭升空而产生的衰减等。但是，如果我们设想火箭经过的距离为y（以英里计），而经历的时间为x（以分钟计）的话，那就不妨假定它们具有$y=x^2$的关系。这个方程可能过于简化了，但是它确实呈现出火箭以加速度运动的本质特征：它意味着，火箭在1分钟后行进了1英里，而在5分钟后行进了25英里。



用微积分方法度量得到的火箭速度

为了计算火箭飞行5分钟时的瞬时速率，我们首先要计算［用莱布尼兹的符号$\left(\frac{\Delta y}{\Delta x}\right)$来表示］在较短的延长时段$\Delta x$中的平均速度。

我们假设$\Delta x=0.01$分钟（比1秒钟还短），这意味着我们的两个时间点（x）分别是5分钟和5.01分钟。利用刚才所假设的$y=x^2$的关系，我们可以计算出，在这一时间段内行进的距离（y）是0.1001英里，从5.01^2中减去5^2即可得到。

现在，我们可以计算这一时间段内的平均速度$\frac{\Delta y}{\Delta x}$，用0.1001（英里）除以0.01（分钟），得到结果是10.01英里每分钟。在更小的时间段内（也就是说，利用极限过程作为微积分基础的基本思想）平均速度就会越来越接近于瞬时速度，即在火箭飞行5分钟（x）这一时刻，速度为每分钟10英里。

实际上我们会发现，在我们假设的“表达式”$y=x^2$的基础上，无论选择哪一个时刻x，瞬时速度都会等于x的二倍。莱布尼兹用他的符号将这一结论表示为$\frac{dy}{dx}=2x$，其中$\frac{dy}{dx}$是$y=x^2$的导数。莱布尼兹将他的符号理解为用无穷小的dy除以无穷小的dx，虽然今天的数

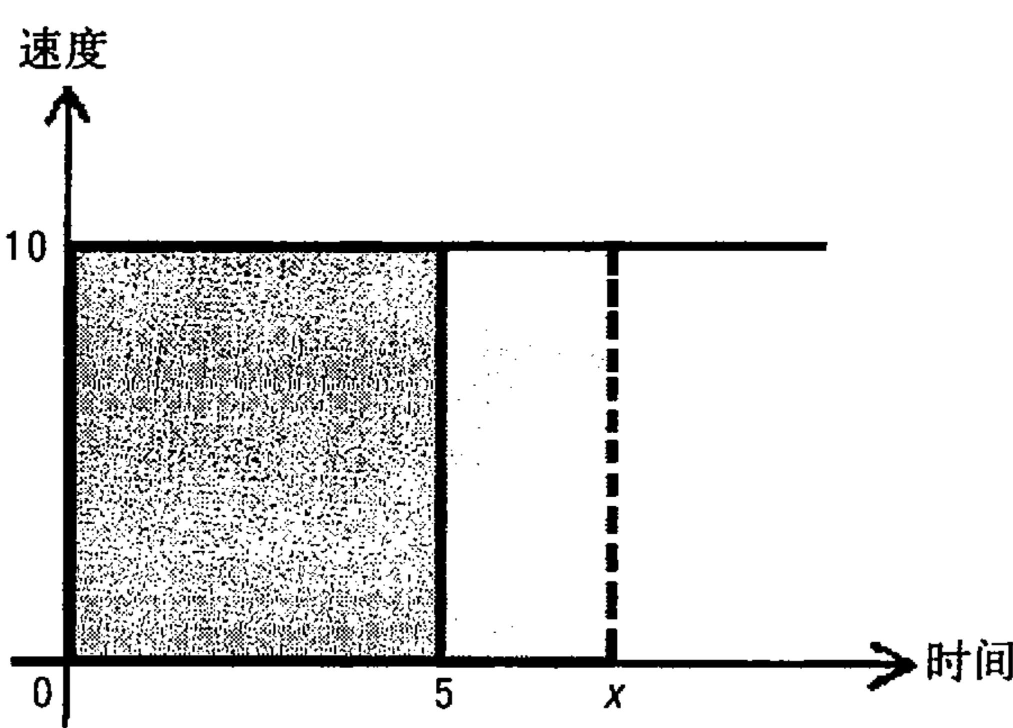
学思维坚持认为$\frac{dy}{dx}$是一个整体的符号。

求导数（“微分”）在微分学中处于核心地位。对于其他数学表达式也可以进行类似的计算，例如，对$y=x$求导得到$\frac{dy}{dx}=1$，而对$y=x^3$求导则得到$\frac{dy}{dx}=3x^2$。

事情的反面

在火箭的例子中，我们假设了火箭行进距离的表达式，接着用微分学知识计算了它的速度。而作为事情的反面，积分学为相反类型的问题提供了解决方法：我们可能已经知道了火箭的速度，而要求的是火箭在特定时刻行进的距离。这依赖于对速度累积效果的计算，数学家们称这一过程为“积分”。

首先我们假定火箭具有恒定的速度（v）10英里每分钟，并且画出速度相对于时间变化的图像，结果刚好是一条经过$v=10$的直线，这表明了速度在任意时刻（沿着x轴即时间轴）都是这个值的事实。

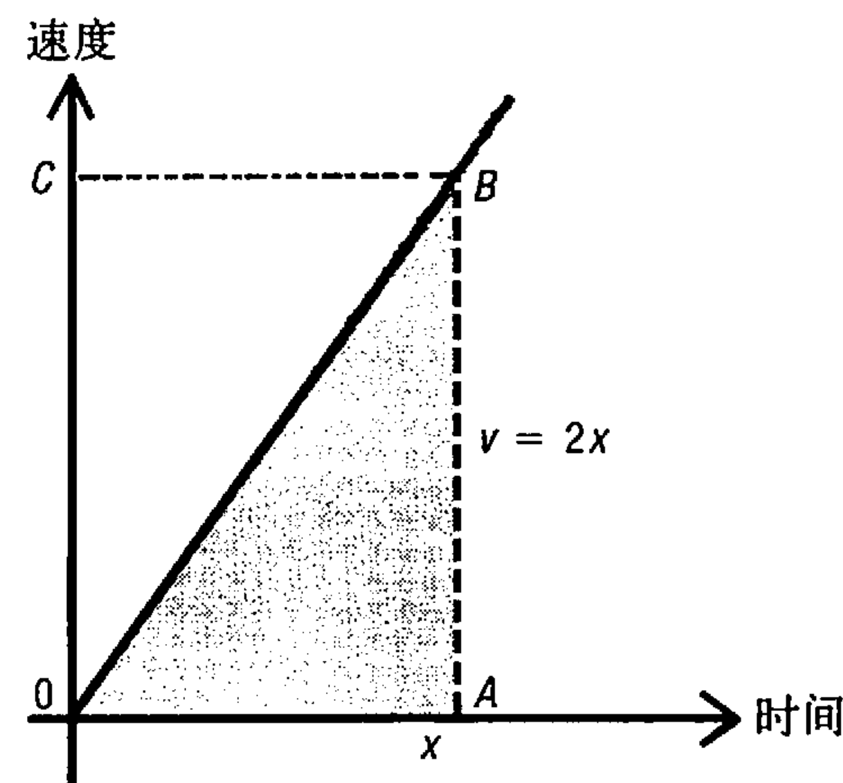


在恒定速度时，用面积表示的行进距离

5分钟后的高度是多少？由于火箭以10英里每分钟的恒定速度行进，因此高度就是5 × 10 = 50英里。从图像表示来看，我们看到距离可以解释为深灰色阴影的面积，它也等于5 × 10。换言之，对于积分学中的逆问题来说，为了求出高度，我们需要计算面积。当然，由于有加速度，我们的火箭并不是以恒定速度行进的。那么，让我们假设速度v是变化的，并且在任意时刻速度（单

位取英里每小时）都等于经过时间量的二倍，即$v=2x$。这意味着，随着时间x的推进，火箭越走越快。现在我们来画与$v=2x$对应的图像，看看结果会怎样。

和刚才一样，火箭行进的距离可以用面积来表示，即图中O, A, B 3个顶点所确定的三角形的面积。如图所示，它是由O, A, B, C 4个顶点所确定的矩形面积的一半。要计算行进距离，先要计算$OA \times OC$，也就是$x \times 2x = 2x^2$。而三角形的面积就是它的一半，即x^2。这就是在假设速度满足$v=2x$的前提下得到的火箭行进距离。



有加速度时，用面积表示的行进距离

计算图像下包围的面积是积分学的核心思想，也是激发牛顿和莱布尼兹的动力。还用这个例子来说明吧，求算图像下面积的过程就是求算小矩形面积$v\mathrm{d}x$之和的过程，沿用历史名词就是“求和”。将它简写成一个字母S，再经过拉长，就形式化为莱布尼兹数学符号系统中的$\int v\mathrm{d}x$，称为v的积分。我们已经发现$\int 2x\mathrm{d}x = x^2$，或者与其等价的（两边同时除以2）$\int x\mathrm{d}x = \frac{x^2}{2}$。

和求导一样，对于其他数学表达式也可以进行积分，例如$\int x^2\mathrm{d}x = \frac{x^3}{3}$和$\int x^3\mathrm{d}x = \frac{x^4}{4}$。

微分—积分的联系与“极限”

那么，怎样将微分学和积分学有效地结合在一起呢？若给定前

例中火箭的高度（y），我们就能用微分学方法计算它的速度——即莱布尼兹的$\frac{dy}{dx}$，而反过来，若给定速度（v），我们也能用积分学方法通过计算$\int v dx$而求出它的高度。用数学语言来讲，这个例子中的关联性质表现为下列等式：

微分学（求速度） 已知$y=x^2$，我们能求出$\frac{dy}{dx}=2x$。

积分学（求行进距离） 已知$v=2x$，我们能求出$\int v dx = x^2$。

微分和积分是互逆的两种运算，牛顿和莱布尼兹都认识到了这一点。（事实上，有时也将积分称为*反导数*，以强调这种互逆关系。）

当然，在牛顿和莱布尼兹刚刚将它发明出来时，微积分还没有完全成形。它是强壮的，但还不够成熟。要为它建立坚固的基础，还需要另一个概念——*极限*。实际上，坚实的基础要到牛顿和莱布尼兹之后很多年才得以形成，甚至有很多顶级大数学家都没能使微积分严格化。到19世纪时，柯西迎难而上，他完善了极限的思想，从而将整个微积分理论建立在更为可靠的基础之上。

微积分与最优化

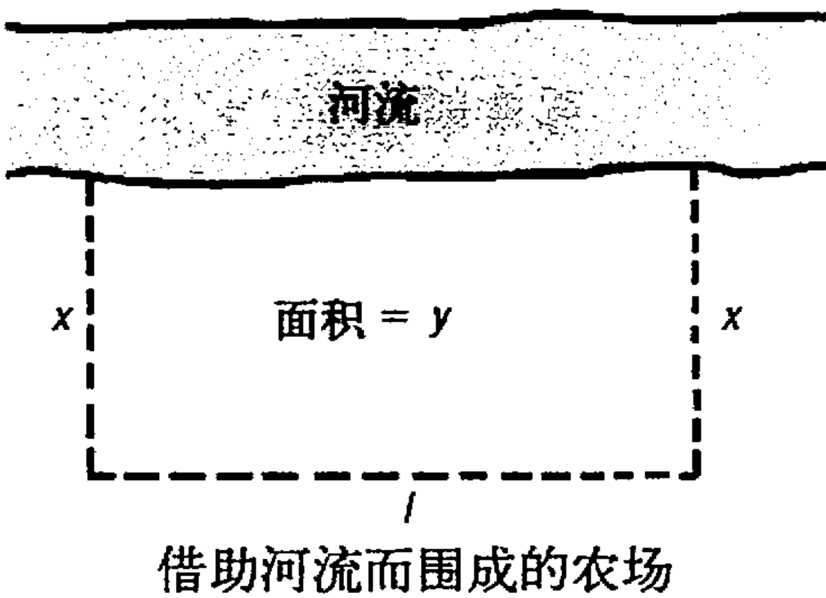
能够证明微积分价值的一项重要应用是可以对某个量求出我们需要的最小值或者最大值。我们可以通过另一个例子来欣赏这种方法，这次我们离开天空，回到大地上。

假设一位农夫有800米长的篱笆，想在土地上围出一块矩形的面积，这块地的一边挨着笔直流淌的河水。他想搞清楚，怎样才能用有限长度的篱笆包围最大面积的土地，以便获得最大的收益。很显然，用河流来充当矩形的一条长边要比充当短边更为有利。接着，

他可能会通过画图，或者在保证总长度800米的前提下尝试各种不同的长度和宽度的方法来试着解决这一难题。但是微积分能够使他省去猜测步骤而直接得到解答。

如果这位农夫将矩形农场的宽度定义为x，而将单独的一个长定义为l，那么矩形的长度l（用米作单位）可以表示为$800-2x$。

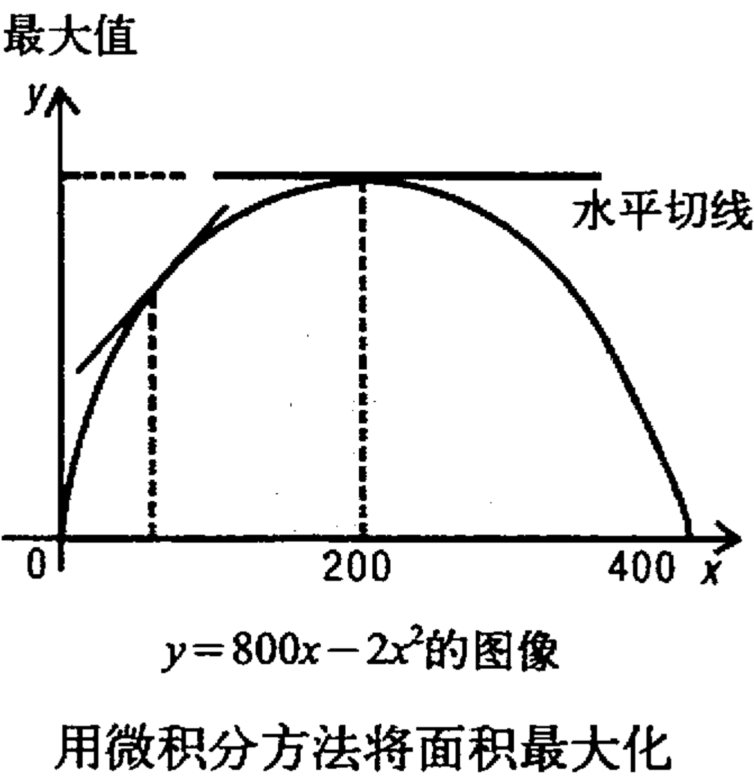
现在，我们可以将农场的面积（它等于$x \times l$）改写为$y=x \times (800-2x)$，进而可以得到$y=800x-2x^2$。例如，当农夫选择$x=100$米时，面积y就是$100 \times (800-200)=60\ 000$平方米。



借助河流而围成的农场

但是，他能否做得更好些，选取一个能得到更大面积农场的x值呢？现在，我们将跳出农夫的世界，把这个问题带进数学家的世界中。通过将上述方程表示在图像中，我们可以看清x取不同值时面积y的不同结果。图中，随着x从零开始增大，面积y逐渐增大，在某一点处取得最大值后又开始减小：图像呈现为一条曲线。

我们在火箭速度的问题中曾经遇到过导数$\frac{dy}{dx}$，但是对于导数还有另一个极富价值的解释，那就是曲线的梯度。曲线上某一点处的斜率是指该点处切线的梯度。这个梯度近似为$\frac{\Delta y}{\Delta x}$，严格地来说等于$\frac{dy}{dx}$。



$y=800x-2x^2$的图像

用微积分方法将面积最大化

在曲线的最高点处（也就是我们想要的那个点处）斜率是0，因

为此处的切线是水平的。因此，如果我们能够计算y的导数，并且令它等于零，就会自然地得到一个关于未知数x的方程。

$800x-2x^2$的导数是$800-4x$，通过求解方程$800-4x=0$，可以得到$x=200$，即为最优解，此时面积为80 000平方米。

从性质方程到数学建模

通常所说的"性质方程"中总有某些项本身就是导数，也就是变化率，因此它们被看做是*微分方程*。举例来说，已知导数$\frac{dy}{dx}=3x^2$，求能够使方程左右两边相等的y。

微分方程的领域十分宽广，除去数学家之外，它还吸引着搞理论研究的物理学家、热衷于化学反应速率的化学家和研究疾病及其快速传播机制的生物学家。这些问题都可以在数学建模的框架之内来进行研究，为了理解一个过程就要先提出简化的若干假设。很多要用到微积分的领域所涉及的量都不止一个变量，例如空间和时间。在这些情况下，数学家们提出了一种叫做*偏导数*的概念，它是通过分别考察每个变量而得到的。它把我们带进*偏微分方程*的王国，这个领域中包括大批有实际应用的经典实例。其中包括纳维尔—斯托克斯（Navier-Stokes）方程（组），它是由19世纪的两位学者建立的，可以应用于气象预测与诸如洋流等流体运动的研究；麦克斯韦方程（组）也出现于19世纪，将电学和磁学联系起来；以及爱因斯坦广义相对论中的场方程。

偏微分方程在经济学中也找到了用武之地，一个典型例子就是布莱克-斯科尔斯方程，它来自美国经济学家布莱克和斯科尔斯，可以用来尝试着预测股票行市。这项工作与关于热分布的经典方程有联系，也使它的提出者获得了1997年的诺贝尔经济学奖。

微积分是牛顿和莱布尼兹留给我们的不朽成就。在科学、社会科学、统计学和工程技术的各个领域之中，很难找到任何一个角落，没有受到这两位伟大的17世纪创造者以及他们对于行星椭圆运动之思考的恩惠。

9 统计学是谎言吗？
——数据，证明与“该死的谎言”

民意调查研究的先驱盖勒普有一段极为著名的言论，说他能够用统计学方法证明上帝存在。持怀疑论调的19世纪历史学家卡莱尔也有一段同样著名的对于统计学的指责，痛斥统计学能证明你所需要的任何事情。这些论调不免有些极端，但确实也引人关注。而不管你是否喜欢它，我们今天都生活在一个统计学无处不在的世界中，大批的统计调研人员忙于收集数据以及利用它们作出推断。但是，他们究竟使用了怎样的数学方法，以及如果可能的话，又能利用统计学得出什么结论呢？

一谈到统计学，人们经常会产生不好的印象，我们带着怀疑的眼光，将它看做是一种似是而非的论证方法，特别是在政治领域中——“统计学”这个词就是由此而产生的[1]。据马克·吐温说，迪斯雷利曾发出“谎言，该死的谎言，以及统计学”这样的指责，尽管那位维多利亚时代的政治家是否真的曾说过这句话是值得怀疑的，但是自那以后，人们便经常利用它来向借助数据“证实”不可靠的论证表示轻蔑。

但是这样的坏名声是否公平呢？不论是否公平，毫无疑问的是，现代社会与经济若没有统计学就无法正常运转：是它在辅佐政府、

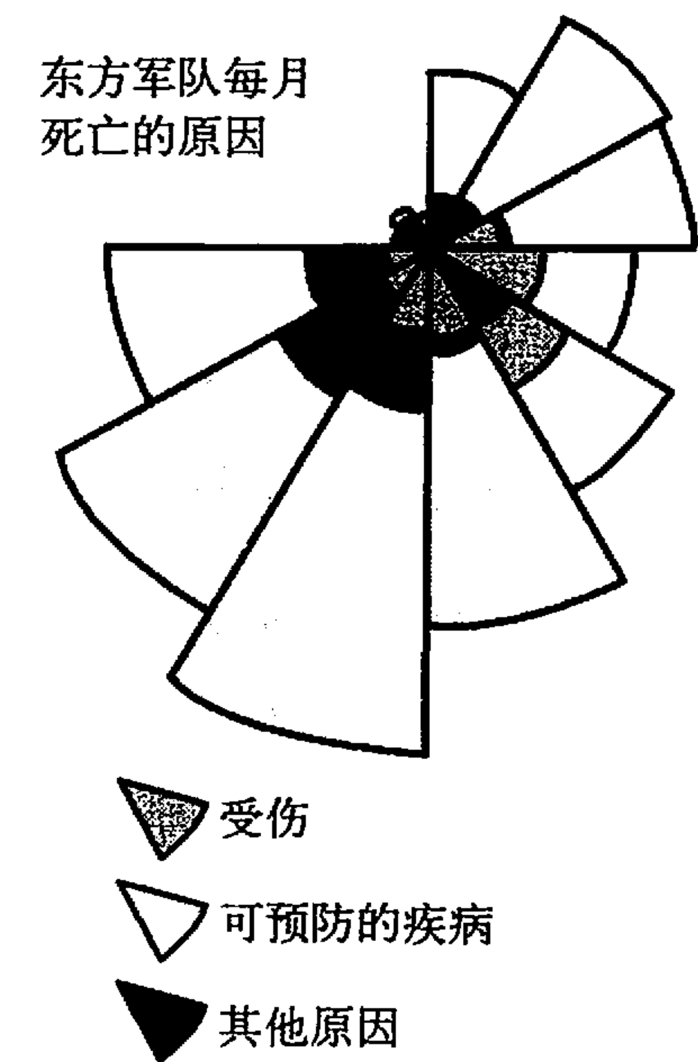
① 统计学（statistics）来源于“state”，即国家、政府的意思。——译者注

国家与国际组织、媒体以及各大型企业做出种种决策。相应地，统计学家也成为供不应求的专家。而且在这个领域中，存在着一些能够体现出数学知识与实际应用结合得非常紧密的实例，它们令人鼓舞而且不是很抽象。

扇形统计图与“光明女士”

统计图表与推动论证向前发展这一目的之间的联系体现出表示方法的重要地位。利用统计数据可以获得大量的图表，它们经常能够将令人望而生畏的海量数据转化成更容易理解也更富于说服力的直观形式。人们在很早以前就明白这个道理了，例如在19世纪50年代，在克里米亚战争中，南丁格尔在管理英国军队的医护工作时就曾使用过这种方法。她将统计数据重新整理，以便强调英国军人所享有的糟糕的医护条件，进而断言有很多人是因为疾病而不是因为作战受伤而丧生的。她设计出一种被她称为“蝙蝠翅膀”的扇形统计图，以便更有效地显示出死亡的原因。

她在图中列出了每月的死亡率，并且将死亡的原因分3种不同颜色分类显示：可预防的疾病、与战争有关的伤亡以及其他原因，各部分的面积与死亡人数成正比。很明显，可预防疾病这一类在每个月中都是最大的部分。图表的目的是说服——“如果没能成功地使公众用耳朵来听信，那就通过眼睛去影响大家”。在这个例子中，统计学实际上使数据更便于阅读，



南丁格尔对于克里米亚战争所设计的“蝙蝠翅膀”统计图

> 在最初的意义上，“统计学”这个词是指治理国家的科学：对于18世纪那些进行政治算术的学者来说，它的功能就是充当中央政府的耳目。
>
> R.A.费歇，统计学家

因为直观表示能够有效地概括数据。南丁格尔的同时代者（斯诺医生）也引入了一种统计表示方法，他将霍乱病例标注在伦敦地图上。由于分布模式变得更为直观，因而有力地支持了他的结论，即疾病是由于公共水泵的不卫生饮用水而产生的，而不是像人们普遍认为的那样，是由“劣质空气”造成的。上面这两个例子都以极为简单的方式应用了统计学知识，阐明了自己的观点。

平均法则

在关于克里米亚战争与伦敦霍乱流行病的这两个案例中，我们从分布和原因的角度来看待数据总体，并将其表示在最终的图表或示意图上。但是，在统计学中经常出现的问题是，如何将长篇累牍的数据提炼成简单也更重要的结果，例如“平均”。可是“平均”究竟是指什么？如果说一个公司内的平均薪金是45 000英镑，它有可能意味着每个人都能拿到这个数目，但是也有可能意味着大多数人拿不到这个数目，比如说，公司总裁的薪金是200 000英镑，它也被放在一起来算平均值。后一种情况也许是更有可能的，不过我们没有办法具体去了解。

绝大多数人在考虑“平均”时，通常会想到“平均数”，将所有读数加在一起，再除以它们的总个数，就可以得到平均数。这时的问题是，一个孤立的偏高数值或偏低数值会产生这样的“平均”：从技术角度来讲它是正确的，可实际上导致了错误的印象。如果将平均理解为“中位数”，也就是将所有数据按照大小顺序排列后的“中间值”，那么平均这个概念就不那么容易受到一个极高/低值的影响。但是，如果将平均理解为众数，也就是全体数据中最常出现的值，

那么我们就能知道这个公司内大多数人能拿到的薪金。

> 我不知道在赫特福德郡这个特别的角落中是否正弥漫着什么格外令人激动的东西，但是在我看来，目前所发生的争执的数目明显高于统计学为我们指点的正常的平均水平。
>
> BRACKNELL女士，引自奥斯卡·王尔德的《认真的重要性》（1895）

在了解平均状态的同时，统计学家经常还要关注各个数据偏离平均的程度。最简单的方法（还用刚才那个例子），是将相对于45 000英镑的每一单个偏离（例如薪金40 000英镑会产生5 000英镑的偏离）累加起来，即可计算出相对于平均的偏离。接着，将所得的全部偏离之和除以读数的个数就得到平均偏离。统计学家格外关注的，是一种更为复杂的计算方法，即“标准偏差”，用希腊字母σ（读作“西格玛”）来表示。它是对偏离的统计，通常可以揭示出隐藏于表象数字之下的另一种类型的真实，例如，如果一群人作为一个整体变得更为富裕，那么他们之间的不平等性可能也会增加。

抽样

在很多情况下，想要收集某一特定环境中的全部数据，既不现实又费力不讨好，例如总人口数：无论在任何时候，一个政府想要了解它的民众的情况，要实施一次人口普查都是很难的。因此，很多统计学研究的一个重要方面是确定一个可靠的样本群体，接着根据其结果作出推断。

作为例子，让我们来想象一下，我们想知道全民平均身高。统计学家所关心的是，如何选择随机的人口样本以避免任何固有偏差。比如说，在那些婴幼儿数目明显高于平均水平的地区附近抽样是不

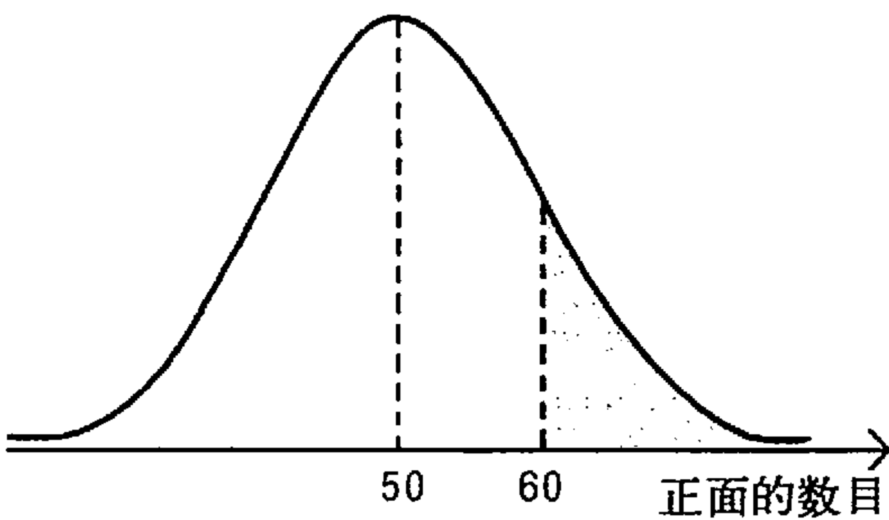
合理的。1936年的美国总统选举为统计学家们提出了随时保持清醒的告诫，预测指出这会是一场难分高下的较量，但实际上罗斯福却以压倒性的优势取胜。在这个案例中，民意调查研究者是借助电话号码本和汽车所有者名单而进行抽样的，因而不小心抽取到一个偏向于相对富足的社会经济群体的样本。

为估计平均身高，统计学家必须求出总平均数，习惯上用字母μ来表示，这个希腊字母读作“miu”。如果所选样本群体是恰当的，那么有可能样本中人口的平均身高［统计学家们用来$\bar{x}$表示它（读作“x拔”）］就是对于总平均数μ的一个很好的估计。

要用严格方式进行估计，就要用到统计学家武器库中的一件重要武器：中心极限定理。它的意思是说，如果重复地进行抽样，那么样本平均数$\bar{x}$就应该服从我们通常所说的“正态分布”。

钟形曲线

“正态分布”用图像来表示就是大名鼎鼎的钟形曲线。它在统计学中的基础地位被人们形容为就像直线对于数学一样。亚伯拉罕·棣莫弗（他的家族在16世纪时从法国迁往伦敦）发现了它的主要性质，不过它获得“正态分布”这个名号还要在经过其他人的努力之后。（有时它也被称为高斯分布，这要源于19世纪时高斯对它的研究。）



抛硬币结果的正态分布

棣莫弗所研究的是一个与传统的投硬币问题相似的问题。如果将一个硬币抛出特定的次数，比如说n次，我们可以计算出得到x个正面的概率。棣莫弗考虑抛硬币次数极大的情况，从而找到一种快

速求出概率的方法，因为此时它们服从正态分布。

如果我们抛硬币100次，那么得到正面的分布的平均数是50。这没有什么可奇怪的，因为这与得到正面的次数大致上占总数一半是一致的。棣莫弗向我们指出，怎样通过计算钟形曲线下的面积来度量概率。例如，得到超过60个正面的概率意味着要计算60右侧的曲线之下的面积。（事实上，统计学家可以借助正态分布表来实现这样的目标。）我们会发现，得到60或更多个正面的概率是0.0228。如果我们想要重复抛100次硬币这一活动，可以期望会在这段时间中有2%的几率得到超过60个正面。

> “某些人厌恶统计学这个名称，但我却觉得它充满了美和乐趣……对于那些从事科学研究的人来说，要想披荆斩棘破路前行，它是唯一可用的工具。”
>
> **弗朗西斯·高尔顿《自然遗传》（1889）**

将抽样调查的实践活动与正态分布联系起来的就是中心极限定理。这个定理告诉我们随机抽样的平均数（即$\bar{x}$）服从以平均数μ为中心的正态分布，并且对于估计μ而言，较大样本所给出的答案会比较小样本更为精确。根据这个定理，如果已知总人口的身高偏差，那么抽取样本平均数$\bar{x}$的偏差就只占它的$\frac{1}{\sqrt{n}}$，因此，随机抽取的样本越大，所得到的总平均数μ的估计就越好。这样，我们就将统计学作为科学来使用了。

统计科学

南丁格尔的图表能够传达出某种信息，并且有效地证明了她的观点，但它并没有使用统计学理论。相对而言，统计科学是指根据

抽取的样本对总体作出推断时所使用的数学理论。

不仅如此，统计学的本质在于要对我们根据理论而作出的预期与我们在实验中实际得到的结果之间的差异进行量化。因此，在抛硬币实验中，我们预期抛100次硬币会得到50次正面，并且可以度量偏离该结果的概率。杰出的英国统计学家费歇在他的《供研究人员使用的统计方法》(1925）中阐述了一种处理这类问题的方法，这是一种实用性的指导，他将其描述为“显著性检验”。

费歇关于显著性检验的著名实例是对于冲茶技艺的研究。究竟应该将茶倒入杯中的牛奶里，还是应该先把茶倒入杯中，然后再加牛奶呢？与他相熟的一位女士宣称她能够分辨出两者之间的差异。作为一个持怀疑论调的人，费歇先提出一个假设：她没有这种能耐。接着，他用品茶实验来检查她的说法。如果她能在8次实验中取得8次成功，就说该结果在统计学意义上讲是显著的，因为纯粹靠碰运气是不大可能得到这一结果的。接着，他就会拒绝那个不大可能的假设，并且承认冲茶确实比他所假设的更复杂。

费歇怎样将显著性检验应用于抛硬币问题呢？我们如何确定一个给定的硬币是否公平呢？也许它的重量分布是不均衡的，但是我们不能进行任何操作来对其进行检验。在进行实验之前，费歇先提出假设“该硬币是公平的”，他将其称为检验“零假设”，这个词来自于物理学家，他们将一次没有得到读数的成功实验称为“零”实验。通常将零假设记为H_0。于是，费歇就能够以高度的信心来接受H_0或者拒绝它。

假设我们得到了一个极端结果：抛硬币100次，得到100个正面。我们能否据此得出结论说该硬币是重量不均衡的，因而是不公平的呢？还不能那么肯定，因为我们可能是偶然地得到100个正面。但是，在假设H_0为正确的前提下，这个结果是极其不可能出现的。事实上，

用一个公平硬币得到100个正面的概率是0.5自乘100次，这是一个非常小的数，大约为8×10^{-31}（也就是在小数点后面有30个零）。于是，我们有充分的理由断言，这个硬币是公平硬币的可能性微乎其微。

> “统计学是一门理应获得敬重的科学，它是很多极其重要之科学的基础……[但是]运用它需要有头脑的人。”
>
> 托马斯·卡莱尔，
> 《批评的杂文集》（1860）

如果结果是60个正面，我们还能得到同样的结论吗？换言之，实验结果偏离预期的数50多少时，才会引起怀疑呢？根据硬币为公平的假设，正面次数为60或者更多的概率大约是2%，要使费歇质疑该硬币是否公平，这个概率是足够低了。

20世纪30年代，内曼和皮尔逊提出了一种足以与之媲美的检验假设的方法，实际上，它导致了与费歇之间的一场激烈论战。在费歇的理论中，只存在着一个假设，关注点集中在是否要拒绝它。而在内曼-皮尔逊理论中，存在着两个*竞争性的*假设。它们的决策模型导致了对于备选假设H_1的考虑。还是用抛硬币的例子吧，内曼和皮尔逊可以假设：

H_0：*硬币是公平的。*

H_1：*硬币是不公平的。*

或者，根据实验的目的不同，他们可以假设：

H_0：*硬币是公平的。*

H_1：*硬币是倾向于正面的。*

在这种方法中，决策规则是在进行一切实验之前就已给出的。我们不再需要等到发现某些不大可能的事情。不仅如此，理论中还引入了两个犯错误的概率。

当硬币实际上为公平（H_0是正确的）时断言硬币为不公平（拒绝H_0）的概率。

当硬币实际上为不公平（H_0是错误的）时判定硬币为公平（接受H_0）的概率。

内曼-皮尔逊理论赢得了信任，因为它致力于使概率更为清晰，还因为它有效地利用了两种犯错误概率之间的“平衡”。

费歇的弟子们继续就这些问题与内曼-皮尔逊方法的追随者展开争论。还有第3条道路，是由“贝叶斯学派”的信徒所倡导的，这个名称来自于18世纪时的牧师和数学家托马斯·贝叶斯。贝叶斯学派想要引入关于一个假设为真的先验概率的思想，并且将结论限制为形如“根据我们已获得的数据该硬币为公平的概率是这样一个数值”的陈述。争论仍在继续。

统计学、证明与真理

无论统计学家们采用哪种竞争性的方法来增强其成果的可靠性，他们的成果都已遍布我们生活的每个角落。政府根据犯罪、移民、就业以及其他很多方面的统计数据来评判其政策。实验心理学家在研究直觉、记忆和注意力持续时间时要用到统计学知识。市场调研人员不断地向我们询问着经济状况和购买选择的问题，而这些回答保证那些想要将商品卖给我们的贸易公司的巨额收入。社会学家和人口统计学家通过抽样方法进行大范围的实验，而诸如热力学和某些力学分支等“硬科学”也一样要从事统计分析。在天文学中，高斯开创了用“最小二乘”技术来将观测误差最小化的道路。这方面研究使他得以更好地分析观测数据，从而成功地在小行星谷神星被太阳遮掩之后重新确定其位置，不过，“最小二乘”技术在现代统计学理论中的价值可远比发现一颗小行星要大得多。

归根结底，统计学究竟能够证明什么？同以往一样，这还是取决于我们所提出的问题，利用统计学方法能够并且已经令人信服地论证了很多事情。不过，统计学并不能证明因果关系。应该说，它所指明的是关联性。回到我们刚才的例子，如果我们最终发现身材较高人群总是住在比较富裕的地区，就可以阐明收入与高度之间的一种关联性，但要继续指出一件事是另一件事的原因，那就步子迈得太大了。

但是在某些情况下，深入的统计学考察确实能够减少甚至解决分歧。在反吸烟运动中，吸烟与癌症之间的因果联系是根本不可能得到决定性的确立的，我们并没有证明癌症出现的生理学原因是由香烟进入肺部而产生的。但是，通过比较吸烟人群与不吸烟的实验对照人群，我们确立了统计学意义上显著的关联性。保守的估计指出，在每年死于肺癌的人中，有80%是烟民，这其实已经足够作为证明了。

尽管某些统计学方面的努力可能会带来惨痛的失败，就像1936年美国大选中所发生的那样，统计学家还是能够得到引人关注的正确结果。现代民意调查是在投票结束前对选举者进行抽样调查而得到的，经常可以是高度精确的。例如2010年的英国大选，被认为是添乱的民意调查，最初曾惹来政治专家的嘲笑，等到实际结果出来，终于为统计学正名。

干巴巴的统计数字本身不能以无限制的确定性证明什么事情，而在不恰当的人手中，它们还可能被滥用。而在人类事务中是几乎不可能实现确定性的，因而我们只能按照以其为真的某种概率去作出决断。通过精心设计的统计调查，我们希望为生活中的不确定性加一道限制。在这个意义上讲，统计学方法绝非空洞无聊之事，更不是谎言。在绝大多数情况下，它是我们所拥有的一切。

10 数学能够保证带来财富吗？——不确定性、机会与概率论

人类总是希望能够减少偶然性。在最广泛和最基本的意义上讲，我们无时无刻不在努力追求对身边环境的掌控力，以及降低风险——包括作物歉收、自然灾害，以及每天潜在的意外事故和不幸。当然，问题也有较为轻松的一面，它反映出人类好玩、好胜的天性，以及人类在生计窘迫之时对于财富的憧憬。数学家们为了减少偶然性而付出种种努力，但现代概率理论兴起得相对较晚，其源头是17世纪时赌徒们进行的机会游戏。但正是骰子和纸牌使一种现代理论得以产生，并且发展成我们今天用来评估风险和应对不确定性的理论的核心部分。

赌博可能使我们变得富有。我们每星期都在目睹这样的事情，当国内及国际彩票抽奖结果揭晓时，百万富翁的队伍中又会增加新的成员。在这个充满各种复杂的经济活动的世界中，五花八门的赌博活动创造出种种机遇，无论是在赛马场上还是在拉斯维加斯的一个放纵周末，漫不经意的一次下注就有可能迎来幸运之神。但是有赢家就要有输家。那些赛前被视为必胜者却最后到达终点的赛马究竟是怎么回事？股市崩溃又是怎么回事？为什么有的赌徒会“连衣服都输光”而早早离开赌场？

多种形式的赌博提出了不同的挑战。在抽奖活动中，你需要的就仅仅是运气，因为过去式不足以指导未来。但这并不适用于股票

交易员，后者需要对公司的记录详加审查。在体育博彩活动中，一位精明的赌徒需要了解该项运动的规则、运动参与者的近况，还要考虑各种偶然因素。而一位拉斯维加斯的度假者则必须懂得见好就收的道理。

> 一切知识最终归结为概率。
>
> 大卫·休谟，《论人的本质》

知识和直觉都可能提供巨大的助力，但是归根结底，是否存在一种合理的打赌方法，使我们在钱财岌岌可危之时做出最佳选择从而避免潜在的危难呢？解决这个问题的第一步，是对于如何计算偶然性建立一种正确的理解。这意味着我们要探讨数学，特别是“概率”的数学。

纸牌和骰子游戏

赌博中最原始也最简单的问题就是抛硬币。如果将一个硬币抛向空中，得到的结果一定是正面（H）或反面（T）。当然，它也有可能以边缘着地，但这种可能性实在太小了，因此我们完全可以在抛投时忽略这种情况，而再抛一次。我们有理由认为，得到正面的机会是两次中就有一次，或者，在技术层面上更为精确地讲“正面的概率是$\frac{1}{2}$”。同理，得到反面的概率也是$\frac{1}{2}$，由于有这一相等关系，我们经常使用的通过在体育比赛前抛硬币来决定参赛队伍或选手的场地分配的方法是公平的。

概率计算奇妙的一面在于它使我们的直觉和预期产生怀疑。我们还是用抛硬币的例子来说明这一点。抛多少次硬币更有可能得到两次正面的结果呢，是3次还是4次？凭借直觉，我们可能会选择第二个答案，因为很明显，它给予我们更多的机会。如果抛3次硬币，那么总共会有$2\times2\times2=8$种可能的结果。得到两次正面的结果有3个，HHT，HTH，THH，因此得到两次正面的概率就是$\frac{3}{8}$。

现在让我们抛4次硬币。在这种情况下，总共有$2\times2\times2\times2=16$种可能的结果，其中包括6个我们需要的结果：HHTT，HTHT，HTTH，THHT，THTH，TTHH。这时得到的概率是$\frac{6}{16}=\frac{3}{8}$，和前面那个一样！抛4次硬币会得到更多的结果，但是概率是有效结果相对结果总数的比率，这就解释了我们刚才得到两个相同答案的事实。

17世纪的业余数学家安托万·贡博曾致力于一个掷骰子游戏，他的问题是，将一个骰子抛4次得到一个六，和将两个骰子抛24次得到一对六，赌哪一个更合算。根据当时的流行看法，第二种情况也就是得到一对六是更有可能的，因为掷骰子的次数要多得多。数学能够解决这个问题吗？

在一次掷骰子后没有得到六的概率是$\frac{5}{6}$，因此掷4次都没有得到六的概率就是：

$$\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}=\left(\frac{5}{6}\right)^4$$

从逻辑上讲，全部的可能性就是要么一个六也没有得到，要么得到至少一个六：毫无疑问，至少有一种情况会出现。一定会发生的事件，其概率是1，因此我们可以说，一个六也没有得到的概率加上至少得到一个六的概率，和为1。这意味着至少得到一个六的概率就是1减去一个六也没有得到的概率：

$$1-\left(\frac{5}{6}\right)^4=0.5177\ \text{（或者近似为51\%）}$$

如果我们将两个骰子一起掷出，那么每次掷出后都有36种可能的结果。这个结果是这么得来的：将第一个骰子上的一点与第二个骰子上的每个点数分别组合起来（得到6个结果），接着，将第一个

骰子上的两点与第二个骰子上的每个点数分别组合（又得到6个结果），以此类推，总共得到36个组合。在这36个组合中有35个都不是一对六，因此没有得到一对六的概率就是$\frac{35}{36}$。这意味着掷24次之后都没有得到一对六的概率是$\left(\frac{35}{36}\right)^{24}$。

通过采取与前面计算一个骰子的情况时相同的过程，我们得到至少有一次能得到一对六的概率是：

$$1-\left(\frac{35}{36}\right)^{24}=0.4914$$（或者近似为49%）

49%比51%小，换言之，掷24次后得到一次一对六的概率略小一些。

彩票狂热

与现代国家或国际彩票抽奖活动中所涉及的数目相比，用一枚硬币或一对骰子所能得到的可能排列数实在微乎其微。有些人会误以为在这种彩票抽奖中获奖的机会和被小行星砸中头部的机会是一样的；其实，通常总是会有一个幸运的获胜者，一位一夜间诞生的百万富翁，而这一事实使我们相信，获胜的可能性还是真实存在的。但是机会究竟有多大呢？

> “我认为，无论你是否参与彩票抽奖，中奖的机会都是一样的。”
>
> **弗兰・里博维茨**

英国全民彩票是一个很好的实例，下注者在1～49范围中选取6个数字，如果它们与机器随机选取的数字完全吻合，那么这位幸运的持票者就能赢得累积奖金。利用概率论，我们可以计算出幸运降临的概率。对于第一个数，共有49种可能性，由于这个数不能与第二个数

相同，因此第二个数有48种可能性。那么前两个数的组合就有49×48种可能性。依次类推到全部6个数，总计可能性的数目为：

$$49 \times 48 \times 47 \times 46 \times 45 \times 44$$

这个计算是考虑到我们所选数字的顺序的，但是在彩票抽奖时我们并不关心它们的顺序。给定6个数，它们共有6×5×4×3×2×1种排列方式，将上述考虑合在一起，得到：

$$\frac{49 \times 48 \times 47 \times 46 \times 45 \times 44}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 13\ 983\ 816$$

这就是6个数所能得到的组合的数量，而且其中只有一个是胜利者。这表明获胜的可能性是很小的，它等于$\frac{1}{13\ 983\ 816}$，近似为0.000 007 15%!

有一种方法可以保证我们一定能中奖。我们可以买足够多的彩票，多到覆盖全部可能的数字组合。若每张彩票1英镑，那就要花13 983 816英镑。但即使如此也还是有风险的，因为我们无法排他性地占有全部数字组合，因此我们可能不得不与其他持有获奖数字组合的持票者分享胜利果实。在赌博的世界中绝对不存在“必然的事情”。

蒙蒂·霍尔谜题

尽管彩票中奖的机会是如此之小，但是关于中奖概率的计算看起来确实还是合乎逻辑的。而某些概率谜题看起来却像是不遵循逻辑的。一个最令人困扰的谜题（无论是对于数学家还是对于非专业人士）当属蒙蒂·霍尔所提出的问题。他在20世纪60年代凭借“Let’s Make a Deal”节目叱咤美国电视荧屏，在这个节目中，参赛者可以通过选择盒子而赢得奖品。

游戏规则是这样的。在3个盖着的盒子（A、B和C）中各有一张纸。在其中两个盒子里，纸上是空白的；而在第3个盒子中，为获奖者提供了一份免费度假旅游的奖品。获胜者要做的就是试试自己的运气，看能否猜中有奖品的盒子。

比如说，我们选择了一个盒子，但是蒙蒂（他了解每个盒子里面藏的是什么）并不是马上让我们打开它。相反，他为我们打开另一个盒子，当然，他会很小心地不打开有奖品的那个盒子，并且允许我们做最后一次选择。于是难题来了：我们应该坚持原来的选择，还是改换成另一个未打开的盒子？

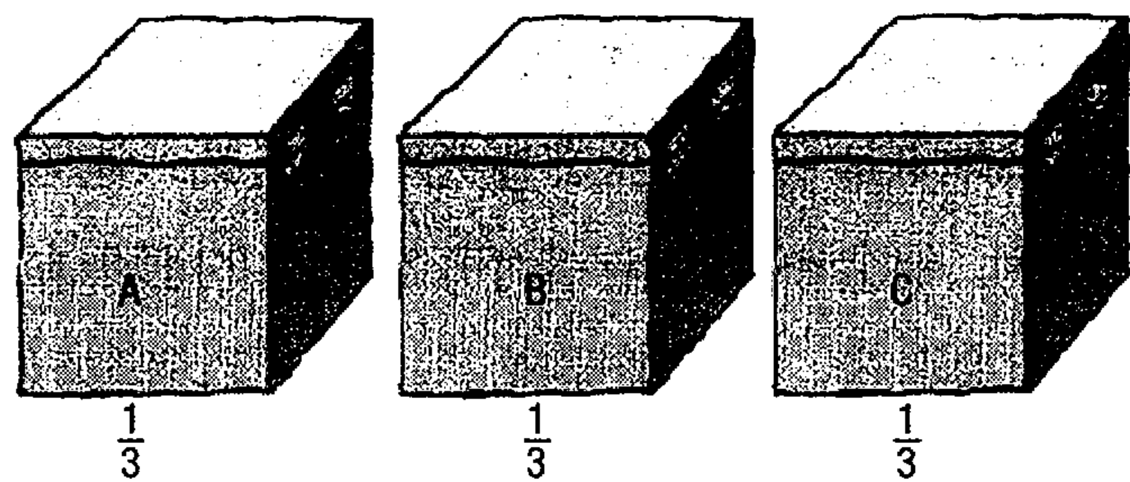
> “在没有能力确定何者为真时，我们应该以最有可能的方式行事……”
>
> 笛卡儿，《方法论》

一个忠告是“不要换”。由于蒙蒂已经告诉我们一个盒子中放的是白纸，所以我们知道剩下两个盒子中有一个里面放的是奖品。无疑，这是一个五五开的选择，那么改换又有何意义呢？从一个角度来看，这个推理是有道理的：我们无法确定地知道哪个盒子里面有奖品，因此我们只能根据哪个选择更有可能这一想法来作出决定。

由此导致的悖论是：不改换选择会减少我们获胜的可能性。推理过程是这样的：有3个盒子，那么选中有奖品的盒子的概率是$\frac{1}{3}$，因为我们没有理由偏好某个盒子。

让我们先假设奖品就在盒子C中（当然，蒙蒂了解这一点），但我们开始时选的是盒子A。蒙蒂当然知道，他要给我们看的只能是盒子B，因为他不可能把有奖品的盒子C打开给我们看。现在我们要做出选择：坚持A还是改换成C。如果改成盒子C，那我们是幸运的。问题在于，如果我们一开始选择了盒子B，那么情况是类似的。在这

种情况下，蒙蒂要给我们看的就是盒子A，而那时我们要做的选择将是坚持B还是改换成C，而如果我们真的换成C，就走运了。因此如果我们改主意的话，中奖的概率可以表示为：$\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$。



蒙蒂·霍尔谜题

相比之下，如果我们一开始就选择了有奖品的盒子C，那么蒙蒂就只能打开盒子A或者B给我们看。这时，如果我们改主意就有$\frac{1}{3}$的概率会丢掉奖品。因此，考虑到所有可能的情况，通过改换选择，我们就能够使中奖的概率$\frac{1}{3}$翻一番，变为$\frac{2}{3}$。问题的关键就在于蒙蒂为我们提供了信息。当盒子的数目减少到2时，我们并不是在一无所知的情况下进行思考的。

扑克玩家与算牌者

今天，一个漫不经心的赌徒已经没有必要再去参加电视台节目，甚至也不需要进入赌场或者彩票销售点。到了21世纪，我们实际上已经处于一个在线扑克游戏的虚拟世界中。就传统而言，作为正牌赌徒的纸牌游戏，扑克牌游戏有很多种名目纷杂的变体，但是最基本的一种，也就是一度盛行于西部蛮荒地区的那种，是“打五张”。先给每位玩家发五张牌，接着做出决定，打出无用的牌并且接受新的牌。

赌徒们根据牌的属性将手中的牌分成若干个等级。最高级的一等是同花大顺（同一花色的A、K、Q、J和10），紧随其后的是同花顺（同一花色中五张相连的牌），以及四张相同点数的牌：如果一个玩家手上的牌比另一个的等级高，那么他就赢了。在双方手上的牌均为最低等级的情况下，胜利将取决于各自手上最大的一张牌，谁大谁赢。

一个理智的扑克牌玩家，或者任何一个打算接受这种挑战方式的人，最好还是考虑一下不同纸牌组合出现的数学概率。从一副52张的扑克牌中取出5张，意味着全部可能的组合数为：

$$\frac{52\times51\times50\times49\times48}{5\times4\times3\times2\times1}=2\ 598\ 960$$

一副牌中只有4个同花大顺。因此拿到一个同花大顺的概率是 $\frac{4}{2\ 598\ 960}=\frac{1}{649\ 740}$ （0.000 15%），实际上是非常小的。赌徒们通常喜欢将得到某种特定结果的可能性用“赔率”来表示，那么在这种情况下，赔率就是649 739∶1。换个方式来讲，不是同花大顺的牌多达649 739倍：这是个有助于保持冷静的想法。而在等级序列的另一端，拿到没有任何关系的五张牌的总概率大约为50%。

当然，抓一手好牌的低概率对于你的对手手中的牌来说也是适用的，和你的牌一样，对手手中的牌很有可能也什么都不是。要在扑克牌游戏中取胜还需要别的本事，因为其中涉及用诈。最终，挑衅式的叫牌以及心理素质等因素往往比牌上的点数更起作用。

但是，数学还能以很多种微妙方式进入赌博游戏。专业的赌徒会用“算牌”技术来估计尚未发出的牌的类型，因此他们能够判断出何时该提高赌注从而提高取胜的概率。事实上，他们用概率论来增加自己的胜算，其中有些人甚至因为取得了巨大的成功而被禁止进入赌场。

理论出场

尽管早期的先驱们就能够运用一种行之有效的概率理论，但是要建立令人满意的理论基础，还要等到20世纪30年代。

这里有一位重要人物，俄罗斯数学家安德烈·柯尔莫哥洛夫。正如希腊人在几何学中强调基础性的公理和公设那样，柯尔莫哥洛夫采用同样的方式为概率论奠定了合理的基础。由于他的工作，这一理论变得抽象起来，也远离了机会的游戏的历史。他的贡献将概率论确立为一门使每个数学家都能有所贡献的统一理论，而不再像一堆彼此无关的想法的大杂烩。

机会游戏可以精确描述出什么是概率。但是真正的挑战在于，如何对一般陈述的事件界定概率的含义。我们如何回答类似“下星期下雨的概率是多少”这样的问题？结果有两种：下雨或者不下雨，但是，与硬币的正反面不同，它们发生的几率是不相等的。数学家们采用的一种方法是为一个主观概率指定一个“置信度”，然后再根据各方面证据对其加以修改。

尽管有这样的困难，概率论还是找到了更为广阔的用武之地。对于科学家，“气体分子运动论”使用概率方式来处理分子运动的问题。在日常生活中，概率是现代风险评估活动中不可或缺的工具，各大企业都要雇用保险精算师，通过计算概率来指定保险规划。在重大法律案件中，陪审团必须要了解如何解释诸如DNA匹配的概率等问题，而在裁定有罪或是无罪时，经常也必须要经常“在考虑概率的前提下”进行。

永远的乐天派

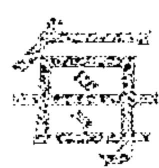
哪里有不确定性，哪里就要用概率论来量度偶然性。要对概率给出一个精确数字，并不总是可以做得到，但是了解概率如何参与

其中，则是一门亟须掌握的知识。从这个意义上讲，机会游戏激励着数学家们将其发展为一门深刻的理论。当然，我们还需要理解它的潜力与极限。再多的概率知识也不能保证让我们在买彩票时一定中奖或者是赢得奖金——但是我们能够了解其中的可能性，从而获得更多的信息，就像参加蒙蒂·霍尔节目的人看到他打开盒子时（不论他们是否真的获得了更多的信息）那样。

正如我们已经看到的那样，一个困难是，数学意义上的概率可能是反直觉的，而在日常生活中，我们中的大多数人在面对需要做出决断的情况时，总是会坚持自己的预感。在玩牌的人和买彩票的人心中，始终还坚持着对于可能性的乐观估计。我们知道数学并不能保证带来财富，但还是要说，也许我们刚好走运……

11 是否存在一个包罗万象的公式?

——数学方法与对知识的探寻

个人都会喜欢捷径。简单来看，数学公式恰好就满足这个要求——前人通过思考而得到的简单方法，使我们不必去重复同样的劳动，而只需将具体数值代入诸如x, y或z等变量即可。但是问题可并不只是这样。寻求公式的渴望是数学和科学中的驱动力。对于科学家来说，公式使理论获得了可信性，从而将知识带入这个世界中。对于数学家来说，公式回答了某个问题，进而建立了一种数学真理。

在常见用法中，“公式”这个词意味着一组模式，一个已经完成的组合结果，以及一种或多或少具有预测能力的方法。“公式”和“方程”这两个词经常是可以互相替换来使用的，尽管严格来讲它们并不是完全相同的事物。对于一个方程，我们的目标是找出使方程左右两边相等的变量的值。公式则是关于一个或多个变量的表达式。

公式的形态和篇幅可谓多种多样。有些具有令人震撼的力量，有些则由于将特定的符号组合起来而使人产生惊奇甚至是美的感受，还有一些则非常有用。最无与伦比的公式当属爱因斯坦的$E=mc^2$，它以如此简明的方式揭示出能量与质量之间的关系，因而使人生出敬畏之心。但是，如果说爱因斯坦的公式是整个节目单中最耀眼的大明星，那么同样从事着重要工作的还有一大批演职人员。

幕后人员

“公式应该是有用的。若非如此，它们也应该是令人惊奇的、精巧的、富于启发性的，或是具有其他某些可以作为补偿的价值。”

安德伍德·杜德利，
《数学杂志》（1983）

在日常生活中真正有用的公式就像老黄牛一样，我们甚至不假思索地就使用它们，正是它们，使我们能够以自动模式进行各种计算。利用我们头脑中的这类公式，我们就可以“按部就班”了——输入一个变量的值，经过简单的计算之后，得出我们需要的那个变量的值。

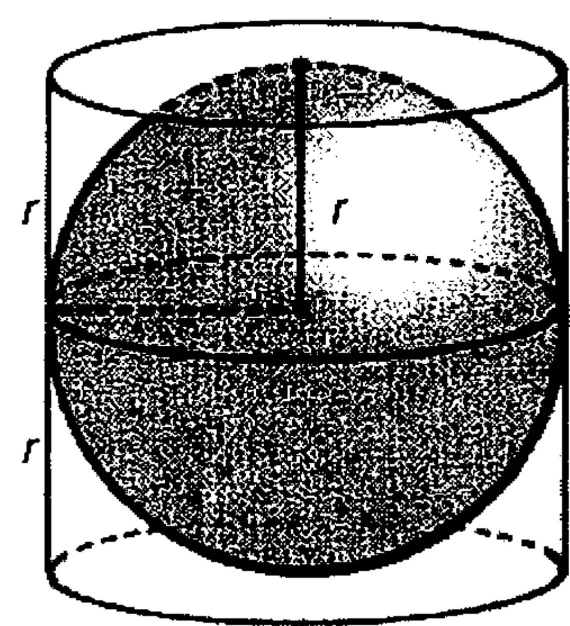
应用得最广泛的两种温标，华氏温标（F）与摄氏或称百分度温标（C），为我们提供了一个直接的实例。它们的名称来自于18世纪时创立该温标的丹尼尔·华伦海特和安德斯·摄尔修斯，尽管前者在科学研究中已不常用，但是天气预报中还在普遍使用它，因此二者之间的转换公式是一种有用的工具。

在华氏温标中，水的凝固点是32度，沸点则是212度，因此这两点之间的温度差是180度。而在摄氏温标中，水的凝固点是0度，沸点则是100度，因此在这种温标中两点之间的差距是100度。由此可知，百分度温标中的一度对应于华氏温标中的$\frac{180}{100}=\frac{9}{5}$度，因此，如果要求出华氏温标下的温度，就要将百分度温标中的温度乘以$\frac{9}{5}$。由于水的凝固点对应于F＝32和C＝0，F开始于32，所以我们的公式应该是：

$$F=\frac{9}{5}\times C+32$$

有些公式构造起来要更唯一些。我们知道一个球的体积取决于它的半径，因为在半径增大时，体积也随之增大。但是这种决

定关系的具体性质是怎样的？一个具有圆形底面的圆柱体的体积是容易求的，方法是用圆形底面的面积（用半径的平方乘以π即可得到，也就是πr^2）乘以圆柱体的高。但是这和球的体积并不是一回事。



球的体积：$\frac{4}{3}\pi r^3$

球体积公式最早是由阿基米德发现的，那是一个灵感迸发的瞬间。阿基米德指出，球体积占将它包围起来的圆柱体体积的三分之二。于是，圆柱体的高度就是底面直径$2r$，而球体积则是$\frac{2}{3}\times\pi r^2\times 2r$，所以球体积公式就是$\frac{4}{3}\pi r^3$。

和其他很多成果一样，这个漂亮的小公式现在已经是积分学中的一个简单练习（参见第8章）。

类似的寻找求长度、面积和体积公式的基本问题经常都是推动数学前进的源泉。古希腊人是研究诸如圆和椭圆等曲线的专家，而我们今天所熟悉的关于这些曲线的公式，是通过17世纪时笛卡儿的代数方法而得到的。尽管我们都熟悉求圆面积和周长的公式，以及求椭圆面积的公式（椭圆面积等于πab，其中a和b分别是两半轴的长度），但求椭圆周长就不那么容易了，在努力理解其性质的过程中，便产生出广阔的数学研究空间，其中包括椭圆积分和椭圆函数等主题。

计算机在现代生活中无处不在，这就产生了对于能计算极大数值的公式的需求。这种公式的背景是计算多个物体全部可能的组合的数学分支——“组合数学”。比如有a、b、c 3个物体，它们可以有$3\times2\times1=6$种排列（即abc, acb, bac, bca, cab, cba），这时候不需要用什么公式。但是，如果我们考虑的是10个物体，那么全部可能的排列数就会多达3 628 800种，要是考虑100个物体，则结果将是

令人瞠目结舌的 $\sim 9.33\times10^{157}$。这时，很明显，我们需要一个能够节省时间的公式。那就是“斯特灵公式”，这个名字来自于苏格兰数学家詹姆斯·斯特灵，它使我们只需计算：

$$\sqrt{2\pi n}\times n^{n}\times \mathrm{e}^{-n}$$

其中，n代表所考虑物体的个数。令人惊讶的是，公式中还包括数学常数π和欧拉常数e（它的值近似为2.718…）。在常见的情况下π是与圆联系在一起的，而e则是与增长等问题联系在一起的，它们的出现提醒我们注意到数学所揭示出来的令人惊讶的联系，特别是因为最初的问题中只涉及整数的乘法。这个公式的另一个引人注目的特点是，作为近似结果，它与真实值相当接近，例如对于100个对象的情况，偏差只有0.083%。

舞台明星

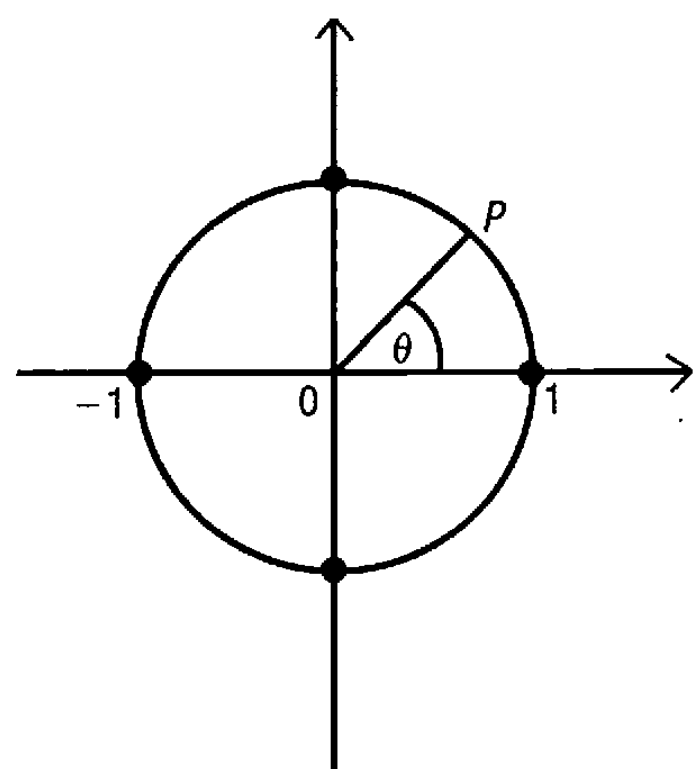
与π和e不同，我们无法将“虚数”i输入计算器中并进行计算，因为它并不是一个实数（参见第5章）。这3个符号都是活跃在数学舞台上的明星，不过若只是看来源的话，它们似乎是毫无关联的。要是再添上0和1，那就是5个大名鼎鼎的数学常数：0, 1, e, π, i。有了这几大明星，数学中最“美丽”的公式当然就是：

$$\mathrm{e}^{\mathrm{i}\pi}+1=0$$

它最早为大家所知是在18世纪，而且与莱昂纳德·欧拉有关系。但是，它是怎么来的呢？要看清这一点，就必须追溯到求“根”的问题。数1有两个平方根（1和−1），我们要将它们表示为一个圆的直径的两个相对端点。这个本身就很“美丽”的圆，把握着那个美丽公式的关键。

接下来我们要温习一下复数，它们可以表示为二维平面中的点

(参见第5章)。一个半径为1的圆可以看做是到原点O的距离是一个单位的全体复数P的集合。可以看到，$e^{i\theta}$是表示这些复数的一种方便的方法，其中θ是从水平轴正向到射线OP的角。



公式$e^{i\pi}+1=0$的几何解释

由于平方根－1表示在该圆上对应于夹角$\theta=180°$（用弧度制来表示就是“π个弧度”）的位置，这意味着$-1=e^{i\pi}$。最后，因为－1＋1＝0，我们就得到了“美丽的”公式$e^{i\pi}+1=0$。

现在我们可以得到一整套漂亮的公式。到现在为止我们考虑了1的平方根，而如果考虑立方根，在下面这个例子中我们就有1的3个立方根，分别处于0°、120°和240°角（用弧度制表示分别为0、$\frac{2\pi}{3}$和$\frac{4\pi}{3}$）的位置，从而得到下面的公式：

$$e^{\frac{2\pi i}{3}}+e^{\frac{4\pi i}{3}}+1=0$$

如果考虑四次方根和五次方根等，就可以得到类似的公式。这意味着我们那个“最美丽的”公式其实只是一长串公式中的第一个而已。

一位丰富多彩的演员

美丽的公式也存在于几何之中，尽管它们不可能总是被用来得到我们想要的答案。数学中一个著名问题出现于19世纪50年代：是否可以只用4种颜色来给地图涂色，就能够保证相邻国家总是具有不同的颜色？用4个彼此毗邻的图形做一个简单的实验，就会发现只用

3种颜色是不可能做到的。看起来对于我们常见的每幅地图4种颜色都是足够用的，但是对于任意地图也会是这样吗？而对于一个数学家来说，“任意地图”不仅包括任何一张已有的地图，还包括任何一张将任意图形置于平面上而形成的地图。这就是后来的“四色猜想”。

没有人能够令人信服地证明这个猜想，不过在19世纪90年代，英国数学家珀西·海伍德解决了一个相关问题。海伍德考虑的是有洞的表面，比如一个“轮胎面”，或者用它更常见的名字——炸面圈形。要多少种颜色才能保证，任意一个画在炸面圈上的地图都可以被正确地着色？海伍德得到了一个非常出色的结果：他给出了一个公式。海伍德指出，对于有r个洞的炸面圈表面，要画出一个正确的地图需要c种颜色：

$$c=\frac{1}{2}\left(7+\sqrt{1+48r}\right)\text{的整数部分}$$

例如，对于形如炸面圈上有一个洞的表面，即当$r=1$时，c是$\frac{1}{2}\left(7+\sqrt{49}\right)$的整数部分，也就是7，因此，用7种颜色就可以在轮胎面上画出任意地图。如果表面上有3个洞，那么$c=9.6033$的整数部分$=9$，因此，用9种颜色就足以保证在这个表面上画出任意地图。

海伍德所得到的公式严格依赖于下面的前提：表面上至少要有一个洞。换言之，当$r=0$时这个公式就不再适用。不过，如果我们真的要这样做，会得到$c=4$这个诱人的结果。不幸的是，一个正常的平面地图恰好就对应于$r=0$的情况，而且我们还必须承认，在制图学领域中，炸面圈形的地图实在是不多。

1976年，在平面上画地图的问题最终得以解决。不过这里不涉及公式，而且，对于在这种情况下是否真的存在一个像海伍德那样

的公式，人们持严重的怀疑态度。结果表明，用4种颜色可以为任意地图着色，只要对1936种关键类型的地图进行检验就可以了。而这件事是由计算机来做的，因此四色定理的证明也成为第一个依赖于机器来完成的数学证明。

助演阵容

我们已经谈到过公式与方程之间的细微差异。但是这又引出了第二个问题：是否总是存在一个解方程的公式？

早在大约公元前2000年时的古巴比伦时代，人们就已熟悉二次方程的各种形式。因此求解二次方程的历史可谓久远。要解一个二次方程，例如$x^2-7x+10=0$，就是要找到使等号左边等于0的x的值。解方程的一般问题可能是困难的，但是对于解二次方程的具体理论来说，存在着一个求根公式：

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

这里，一般形式二次方程中$ax^2+bx+c=0$中的字母a, b, c可以是任意数。这个广为人知的公式只涉及加、减、乘、除和求平方根等几种基本运算。

文艺复兴时期的数学家们将注意力转向三次方程和四次方程，也就是首项分别为x^3和x^4的方程。对于这两种情况，都已找到了求根公式。尽管四次方程的求根公式可谓冗长而繁琐，但是它还是只涉及前述几种基本运算。

现在，对于次数为5的方程，也就是首项为x^5的方程，比起寻找求根公式来说，更为自然的问题应该是什么？这次可真是说起来容易做起来难。数学家们花了300年时间去寻找它，却始终一无所获。这个问题也成为数学中最重要的未解决难题之一。能够给出近似解

的方法是可能的，但是数学家想要的是一个能够给出精确解的公式。在19世纪20年代，年轻的挪威人尼尔斯·阿贝尔终结了（几乎）所有寻找求根公式的努力，他发现，适用于所有五次方程的公式是不可能存在的。例如，没有一个公式能够解$x^5-2x-1=0$。对于这个结果，曾有几位勇者为之而奋斗，而阿贝尔那里程碑式的成果在极大程度上解决了这个问题。

璀璨群星

求根公式对于纯数学家来说是重要的，对于正在校园中学习的很多非数学家来说可能也会引起共鸣。但是要展示出公式的真正魅力，我们还要将关注点从数学转向科学的世界中。的确，长期以来，数学为科学作出的主要贡献就体现在对公式的关注上。

> “可能再没有别的自然定律能像牛顿在《自然哲学之数学原理》中提出的万有引力定律那样，以如此简单的方式将如此多的自然现象统一起来。”
>
> E. T. 贝尔，《数学大师》

例如，伽利略将注意力从关注“一粒石子为什么要落向地面”这样的问题转移到关注“找到石子下落距离的公式”这样的定量问题。这个公式会依赖于哪个或者哪几个变量呢？伽利略的天才在于他发现下落小球的速度并不像亚里士多德曾以为的那样，依赖于小球的质量。较重物体并不会比较轻物体下落得更快。他通过实验发现，经过x秒之后，下落距离d由公式$d=16x^2$所确定。（公式中的系数为16，是由测量时所选择的物理单位决定的，其中测量时间用的单位是秒，测量距离用的是英尺。）伽利略所取得的突破性成果是，用一个简单的公式就可以描述运动并且给出答案。利用这个公式，

我们就可以计算出在未来任一时刻x时石子下落的距离。

牛顿引力理论的核心成果称得上是最伟大的公式之一。要是没有它，我们就无法实现载人登月、太空旅行，甚至不能算是开始理解物理世界。

牛顿理论认为，任意两个质量记为m_1和m_2的物体都会彼此吸引，这个吸引力正比于m_1和m_2的乘积除以d^2所得的商（d是两个质量之间的距离）。这个著名的公式是：

$$F = G\frac{m_1 m_2}{d^2}$$

它被称为“万有引力的反比平方律”，其中G是引力常数，它的数值取决于测量时所选取的单位。我们说牛顿的公式是一种“反比平方定律”，是因为要除以d^2，因此两个物体之间的距离越大，它们之间的引力就越弱。对于行星来说，m_1和m_2非常大，所以它们彼此吸引的力也就会非常显著。

伽利略和牛顿的成功暗示我们，宇宙的运转就像一台机械钟一样，是充分确定的。这是一种强有力的观念，它也为那些试图从观测研究中寻找公式的科学家们提供了动力，如果足够幸运的话，他们的成果也有可能会变成流芳百世的科学定律。在18世纪晚期和19世纪早期，皮埃尔·西蒙·拉普拉斯就是这种观念的一位重要拥护者。他宣称，要确定未来所需要的一切就是广博的才智，以便将某一时刻的原子运动、力、行星的运动，以及一切自然数据汇总在一起——接着将它们输入到一个可以完全确定世界未来状态的公式中！于是，一种确定性宇宙的信念就在对于一个万能公式的美好幻想的基础上建立起来。

“相对论出现以后，物理学家不再需要一个能够从外部审视整个宇宙的魔鬼，但是他们仍然可以想象有一位绝顶精明的数学家，就像爱因斯坦所认为的那样，他既不会说谎也不会掷骰子。这个数学家可以拥有关于整个宇宙的公式，其中也包括对于自然的完整描述。”

埃利亚·普利高津，《混沌中的秩序》(1984)

19世纪时，迈克尔·法拉第通过实验得知，磁和电具有密切的联系。在将条形磁铁穿过导线圈时，会有电流通过导线，而且反过来，在有电流通过导线圈时，处于线圈中的一段铁棒会变成磁铁。这种联系是构成电动机的基础，但是能够将它概括为一个公式吗？这就为数学物理学家詹姆斯·克拉克·麦克斯韦的才华找到了用武之地：他不仅写下“麦克斯韦方程（组）”来描述这种现象，而且还创造出一些新的数学成果。这4个将电学与磁学结合起来的公式是极其重要的。

用来描述物理事件的公式可以简短到令人吃惊，例如波义耳定律（$P=\frac{k}{V}$），它告诉我们一种气体的压强和体积互为反比例关系。在物理学中犹如路标一样的公式，爱因斯坦的$E=mc^2$，是相对论中的一个成果，它所呈现出的令人痴迷的简明性甚至会掩盖住它那深刻的重要性。还有一些依赖于多个变量的公式，比如制造透镜的公式，它揭示出透镜焦距与折射率、曲率和厚度之间的关系。

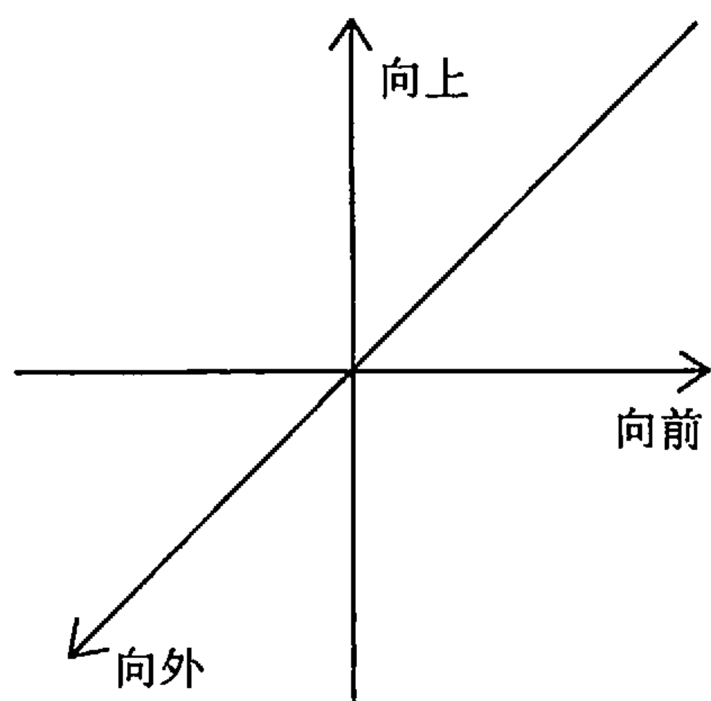
魔法公式？

无论是在数学还是在科学中，一个包罗万象的魔法公式的存在都是虚无缥缈的。比如在数学中，我们知道，不存在可以解所有五次方程的代数公式。事实上，库尔特·哥德尔在20世纪30年代给出

的“不完备性定理”告诉我们，在形式化的数学系统中，存在着真实但却不可判定的陈述，也就是说，它排除了能够自动创造出所有定理的简单公式或者“机械装置”。但是渴求的步伐是无休无止的，数学家们仍在不懈地寻找着模式和公式。科学家在观测数据时，也仍在探寻着潜在的基本系统——这是一切成功的公式的先决条件。今天的科学家们，尽管已经不那么受决定论的支配，却还是渴望新公式的发现。只有这样，才有可能成为科学定律，从而使他们的理论获得可靠性——也使他们因成功而获得回报，在历史的长河中占据明确的地位。

12 为什么三维还不够用？——更高的维度、怪物曲线与分形

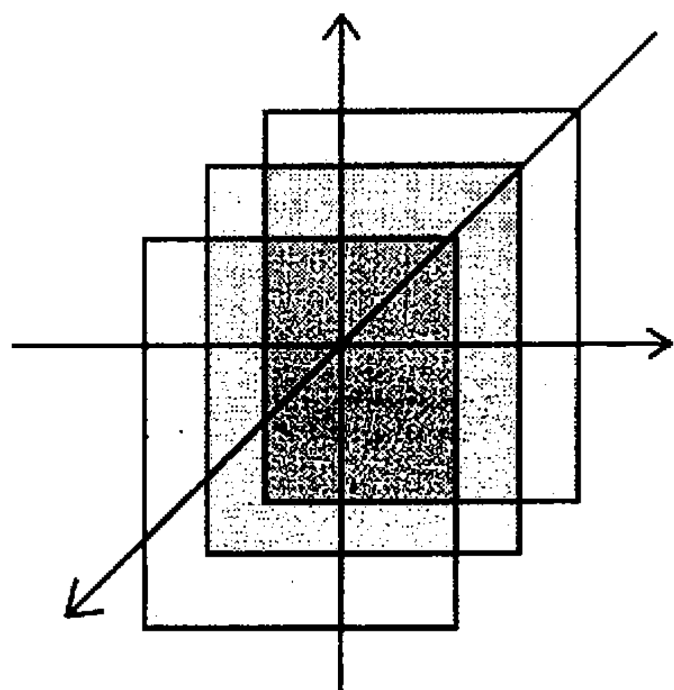
很多个世纪以来，我们只使用三个维度，并且这样就够用了——没人觉得会有问题。但是后来，物理学家宣称，要想真正恰当地理解我们的世界乃至整个宇宙，就需要增加维数。再近些时候，计算机的强大绘图能力将我们带入了一个隐秘的世界，在这里维度甚至不一定要是整数。我们对于维度的全面认识经历了从熟悉到神秘的历程。



传统的三个维度

将我们居处其间的物理世界看做是三维世界，这可以说是老生常谈了。无论身在何处，我们都可以向前、向上和向外三个方向移动——或者是三者的组合。在我们的文化中，三维概念也一直对我们的语言和想象力施加着强烈的影响，如3D图形，3D地图和3D电影。

古希腊人通过一种等级分明的方式来考察维度。点是几何构造的最基本要素，被认为是“零维的”。点汇集起来构成了线，因此线是一维的。线汇集起来构成了面（面是二维的），而最后，连续不断的面再构成三维的空间。这种分等级的概念中暗含着连续运动的思想：点动成线，线扫出面，面面结合成空间。


前后相继的平面形成三维空间

这种关于维度的基本观念是适于数学家和科学家进行探索的。亚里士多德在《物理学》中指出，空间有六个维度：他认为引入反方向的运动是非常自然的，因此向前、向上和向外三个方向就变成前/后、上/下和内/外三个方向对。但是到了文艺复兴时期，莱昂纳多·达芬奇抛弃了六维概念，宣称绘画的科学始于三维几何学。追随古希腊几何学家的足迹，他主张点是最基本的，其次是线，第三是面，第四是空间，空间其实是平面的“叠盖”。莱昂纳多理解他的几何学，并且认为他的学生们理当也是如此。

一个多世纪以后，勒内·笛卡儿改造了几何学，将其转化为代数。利用后来被称为“笛卡儿坐标轴”的x, y, z轴，空间中的一个点可以用三维坐标(x, y, z)来描述。于是，以点$(3, 1, 20)$为例，它就表示前行3个单位，上行1个单位，再向外20个单位后所到达的位置[①]。这种从几何到代数的转换，意味着几何对象可以用方程来描述和研究。例如，$x+y+z=0$是一个平面的方程，而$x^2+y^2+z^2=1$则是以1为半径的球面的方程。它们都是三维空间中的几何对象，不过实际上平面和球面（一个球的表面）自身还是二维的。如果你居处其间或者是到那里去旅行，那么还是会认为自己生活在一个二维世界中。利用代数方法可以获得更高程度的精确性，几何学从此不再只依赖于“看见”一个几何图案或者图形。代数符号还能帮助数学家注意到那些用其他方法可能尚未发现过的几何问题。

① 作者这里所采用的“向前”、“向上”、“向外”的说法，分别对应于我们熟悉的x、y、z轴的正方向。即xoy平面处于纸面内，z轴垂直于纸面。——译者注

通向更高维度之旅

一开始，人们将较高空间维度（人们认为这是不寻常的）称为“超几何学”（类似的还有“超平面”和“超球面”），但实际上，在方程中引入更高维度是相当容易的。对于四维情况，我们可向已有的三个变量中再添加一个，而对于五维则添加两个变量。于是，$x+y+z+w=0$就是四维空间中一个超平面的方程。而$x+y+z+w+u=0$就是五维空间中一个超平面的方程。类似地，$x^2+y^2+z^2+w^2=1$是四维空间中一个超球面的方程，而$x^2+y^2+z^2+w^2+u^2=1$则是五维超球面（参见第17章）。

我们可以考虑的维度数目是没有限制的。通过提出“遗传空间”的思想，牛津进化论生物学家理查德·道金斯建立了一个计算机模型，它呈现出从以前存在的动物直到未来的动物之间所经过的历程。他需要数以千计的维度，每个维度是他所考虑的动物的一种独立基因。

一些数学家在新观念刚刚引入时就对其情有独钟。亚瑟·凯莱早在19世纪40年代就写出了关于n维空间（n代表任意数）的论文，因此只要他自己愿意，没有什么能阻止他去考虑105维或是10 500维。不过实际上，数学家们也不情愿抛弃常用的空间观念，我们居处其间的物理空间始终是三维的，而更高维度的概念只对抽象的数学空间才有用。

> “……你需要恰好三个维度才能打一个结，那种能够拉紧而且不会松开的结，而这正是基本粒子的具有的形态——时空中的打结。我们不可能在二维空间中打结，因为二维空间中没有上下之分……”
>
> 约翰·厄普代克（1986）

但是，我们真的可以一直坚持物理空间就是三维的吗？爱因斯坦提出四个维度才是正确的，因为与牛顿认为时间和空间彼此独立不同，爱因斯坦所考虑的是时间—空间连续统，其中时间和空间是互相影响的。

二维国，只有两个维度的世界

要想真正理解在描画四维世界时所遇到的困难，最好的方法莫过于去拜访“二维国”，那是19世纪80年代的一位教师艾伯特在他那讽刺当时社会的小说《二维国：通向更高维度的浪漫之旅》[①]中虚构出来的一个幻想世界。为了构建他的世界，艾伯特将三个空间坐标缩减到只剩两个。在二维国里，人们只能在没有高度的平坦表面内过日子。他们无法脱离，特别是无法获得对所处世界的总体印象。在二维国里没有“上方”这个概念。

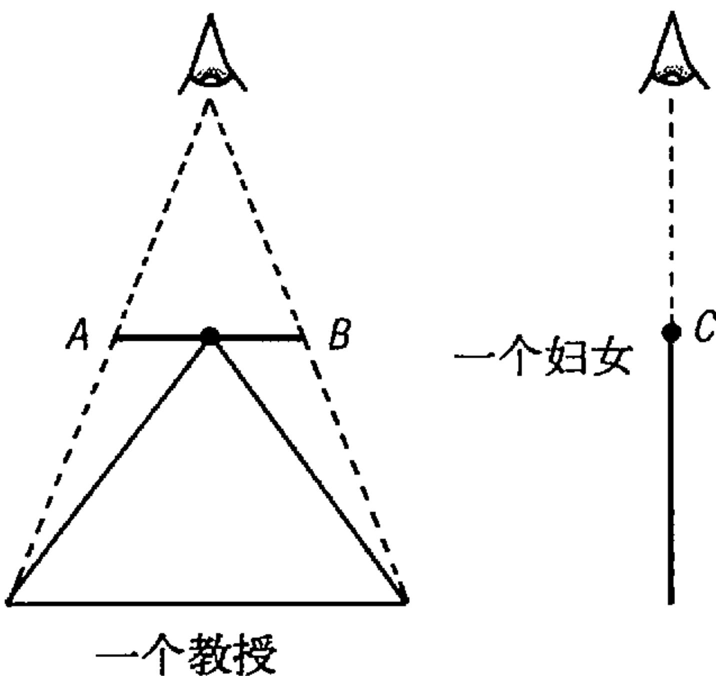
二维国中的一切人和物体都只有两个维度，他们可以在故事里的虚构世界中自由行动，但却无法脱离纸面。人们的形象是各种图形，例如三角形、正方形、五边形以及其他一些对称图形，而且这些形状反映了严格的等级结构。边数较少的图形，例如三角形，处于社会等级的较低端，而且随着图形边数的增加，等级也逐渐升高。由于二维国中的每个人都生活在平面中，不同的形状很难彼此区分。因此每个居民都要找到一些巧妙的方法来互相识别。

以大学教授为例，他是一个等边三角形。因为二维国的所有居民都生活在二维世界中，所以他的学生无法分辨出三角形上的点：他们所能看到的一切就是一条线段AB——当然，除非有光影效果辅助他们的视觉。妇女在二维国社会等级制度中处于最低阶层。她们

① 我国曾引入这本书，曾用译名为意译的《神奇的二维国》。这里选择将原书名直译。——译者注

的形象是“直线”，妇女没有任何办法去赢得社会地位，因为她无法变成一个三角形以提高社会阶层。但是妇女也有一个特别之处。作为直线，她们能够变得几乎看不见：如果迎面看过去，就只能看到单独一个点。

除去讽刺社会的意味之外，我们在二维国中体验到的认识和观念层面的种种困难，与试图想象四维世界时所遇到的困难是完全一样的，区别只是又多出一维而已。我们怎样才能“看见”一个超平面或是超球面呢？

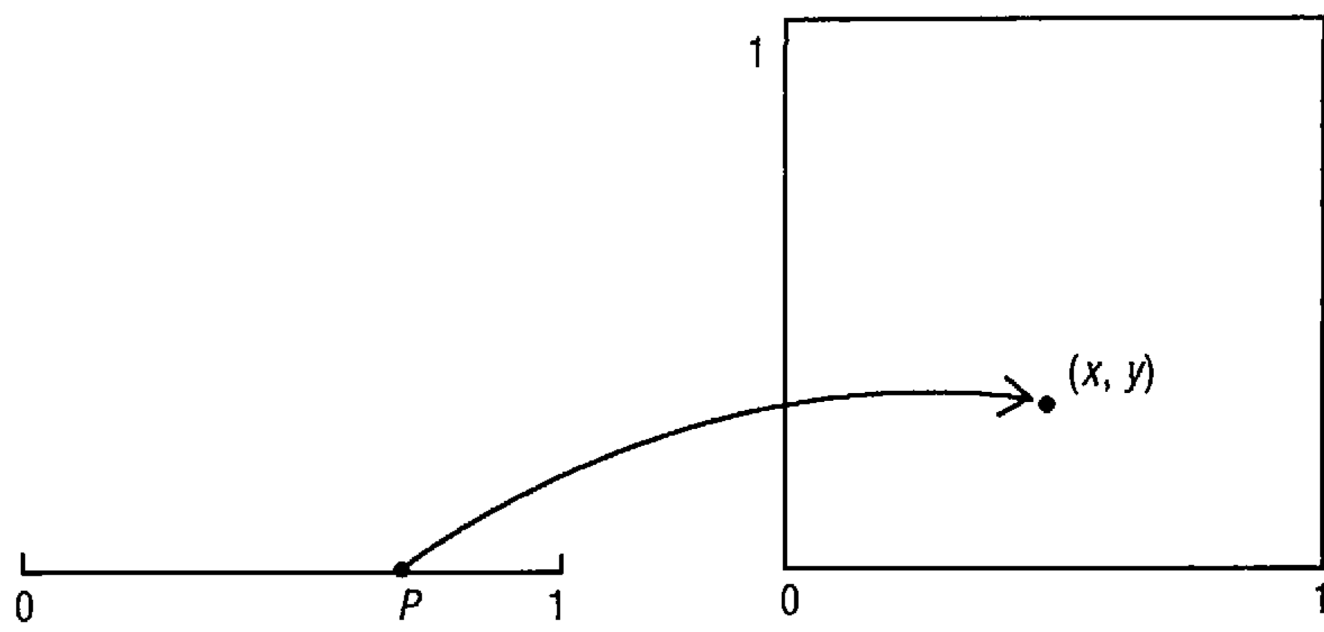


二维国中的二维人物，一个教授和一个妇女

康托尔的惊讶

如果19世纪70年代的几何学家们认为自己已经了解了维度的概念，那他们可能就要大吃一惊了。人们一直凭直觉方式来把握维度这个概念，但从未曾给它下一个严谨的定义。当时在德国哈雷大学从事研究工作的格奥尔格・康托尔得到了一个令人震惊的发现。从传统意义上来讲，我们将正方形看做是二维的，是基于这样一种观念：一个单位正方形是由具有单位长度的线段层层堆叠而形成的。如果一条单位长度的一维的线段是由大量可区分的点组成，那就有理由认为，由很多条线组成的正方形中应该包含着更多的点。

康托尔反驳了这一观念。他在组成单位线段的所有点与单位正方形中的所有点之间建立了一种一一对应关系。换言之，他断言线段上的每一个点在正方形中都有一个对应点，反之亦然。



康托尔所发现的线与正方形的一一对应

康托尔是怎么做到的呢？首先，他将线段上的点表示为一个小数——例如，这个点可能是0.197 625 43…接着，他在正方形内构造出一个坐标为(x, y)的对应点，方法是将原小数中各位数字交替分配给x和y，于是在这个例子中，我们有$x=0.1724\cdots$和$y=0.9653\cdots$反过来说，如果给我们一个正方形中的点(x, y)，那么就可以将x和y中的数字插接起来，构成一个与线段上某点对应的小数表示。

康托尔的结果完全违背了直觉，甚至也包括他自己的直觉。1877年，他在写给理查德·戴德金的信中是这样评论的："我看到了，但是我不相信。"那时，几何学家把一维物体和二维物体看做是完全不同的，但是康托尔迈出了勇敢的一步，通过建立一一对应关系，他断言它们包含同样多的点。他甚至能够在一条线段与一个三维立方体之间建立一一对应关系（利用类似的方法，每隔两个数字一组），而且可以更进一步地与高维立方体建立一一对应。

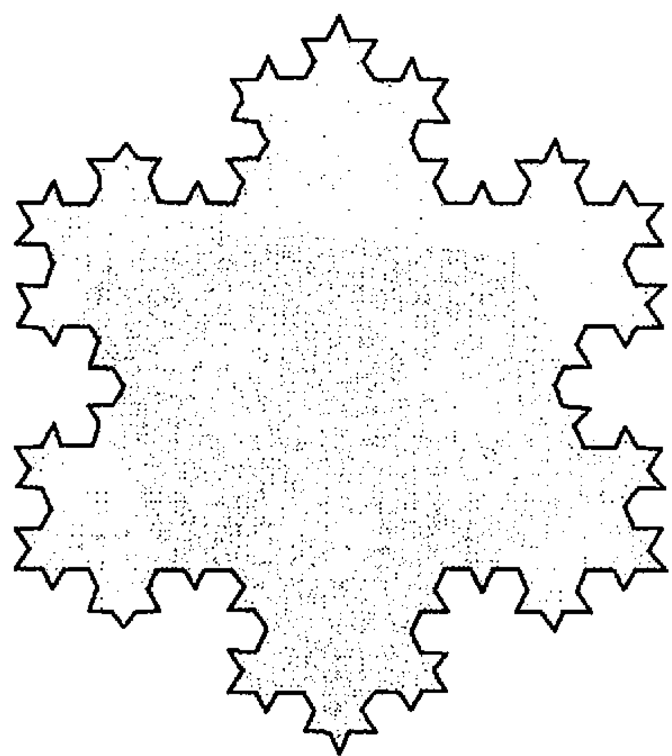
对于数学家来说，康托尔建立的这种一一对应关系中缺少了一种性质。它不是连续对应，也就是说，康托尔不能保证，线段上一个选定点附近的各点全都变成单位正方形中对应点附近的点。25年后，荷兰哲学家和数学家路易岑·布劳威尔证明，不可能通过连续变换将一个物体转化为具有不同维度的另一个物体。

曲线和“怪物”曲线

与维度定义和维度的数目纠缠在一起的是一个经久不息的问题：什么是曲线？古希腊人就已经和圆、椭圆、双曲线、螺线以及其他一些特殊曲线打交道了。但是，那些不属于传统成员的曲线又该如何呢？19世纪80年代，卡米尔·若尔当尝试性地以连续运动的点的方式提出了曲线的定义。

这种定义的好处是，一些相当复杂的图形，比如无断裂的“指纹”，也都可以看做是曲线。麻烦在于得到它承认的对象实在是太多了，根据这一定义，一个实心的正方形也可以看做是一条曲线。一条“一维”曲线如何会变成一个实心的“二维”物体呢？这是荒谬的。为了找到更合适的定义，数学家们开始制造具有各种最奇特性质的“曲线”，这些意料之外的结果与通常情况下所遇到的各种曲线截然不同。比较恰当的叫法是称它们为“怪物”曲线。对于曲线的研究产生出一系列不同于以往但又有所关联的维度概念。

一种出现于19世纪末的怪物曲线叫做科赫曲线，它的名字来自于瑞典数学家赫尔奇·冯·科赫。“怪物”这个词所形容的是它的性质，可不是它的外观——因为它就是随处可见的雪花曲线。要得到这种曲线，让我们先从一个等边三角形（将其定义为“第0步”）开始，接着在每条边上增加一个三角形“突起”，不断地重复该过程，得到一个n阶段序列，雪花曲线就是这个序列的极限。也就是说，通过不断重复可以产生该曲线。我们不可能画出真正的雪花曲线，但是可以复制它的某一阶段。



雪花曲线的一个阶段

随着增加的突起越来越多，雪花的边界也逐渐增大，事实上，雪花曲线的最终长度是无穷大。因为整个图形总是保持在一个圆内，也就是具有有限的面积，所以，它是一条具有有限面积和无穷边界的曲线。

但是雪花曲线的维度是指什么？我们该怎样度量它？在诸多关于维度的定义之中，哪一个是有效的从而应该拿来使用呢？这时，就可以看出德国数学家菲利克斯·豪斯道夫的思想的用处了。就目前所知而言，“豪斯道夫维度”对于正常图形给出了和“常用”名称系统相一致的结果，所以一条线的维度是1，正方形的维度是2，而立方体则是3。这是因为豪斯道夫维度（d）取决于长度和面积的测量。如果将一个正方形的每条边都同样放大为原来的3倍，那么新得到的面积将是原数值的9倍。由于$9=3^2$，所以一个实心正方形的豪斯道夫维度就是其中的指数，即$d=2$。当然，对于正方形来说这正是我们所期待的结果。豪斯道夫维度的关键就是指数。

雪花曲线的生成元素是其中的一段，如果将它按比例放大到原来的3倍，引入一个突起部分，我们会得到一段长度是原来4倍的曲线。因此我们不得不根据$4=3^d$来计算d。d的数值一定介于1和2之间，因为4介于3^1和3^2之间，事实上，$d=1.292\ 24\cdots$

这可是个新事物。我们得到了一条维度不是整数的曲线——它具有分数维度。事实上，我们已经进入了在五十多年后被称为“分形”的曲线的世界。

分形与曼德尔布罗特集

如果我们用放大镜来观察科赫曲线，就会发现一些神奇的东西。在放大镜下，我们看到的是与没用放大镜时一模一样的图形。就算再继续放大，也还是会看到同样的东西。科赫曲线具有自相似性质，

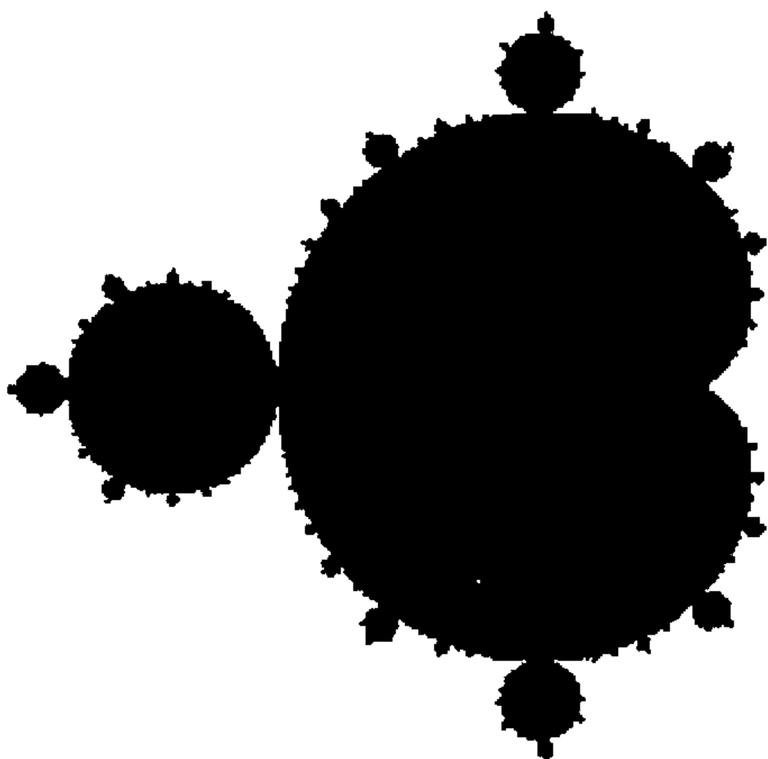
这是由于不断重复的生成方式所导致的结果。类似圆等正常曲线是不会有这种问题的。如果我们将圆周的一部分放大，将会看到一段显得直一些的曲线，而且越是近距离观察看到的部分就越直——但是不会看到很多个小圆。

我们要感谢“分形之父”——数学家本诺特·曼德尔布罗特，是他将分形科学带进了现代意识。分形是一类断裂或弯折的曲线或图形，具有自相似性质，以及按照豪斯道夫维度方式度量的维度。科赫曲线是利用具有正常维度1的直线段构造出来的，但是科赫曲线却具有豪斯道夫维度$d=1.292\ 24\cdots$因此它是一种典型的分形，其维度超出正常维度的范围。

最著名的分形当属曼德尔布罗特集。它是曼德尔布罗特于1980年在计算实验室中发现的，而且它甚至比科赫曲线更奇特。当曼德尔布罗特第一次在自己的行式打印机上看到它时，就为它的错综复杂而感到震惊。这绝对是出人意料的，而且由于它是在很基本的条件下出现的，后来它成为所有分形中的标志。曼德尔布罗特发现了这个集合，它不是像科赫曲线那样通过反复生成突起而得到，而是通过对一个很简单的代数公式的迭代得出的。

曼德尔布罗特集也是自相似的。如果我们将其中一部分放大，就会看到数以百万计的微型的曼德尔布罗特集。而且，如果我们继续将其中每一个放大，就能看到更多个小的曼德尔布罗特集。可是更多的惊喜还在后面，让我们来看看它的维度。曼德尔布罗特的边界和正方形完全不同，但是它的豪斯道夫维度却等于2，和正方形的豪斯道夫维度相同，再一次向我们揭示出维度本身的奇异性质。

当分形为人所熟知以后，它们就拓展了数学和计算机建模的疆域。我们没法用欧氏几何学中的任何图形来对云的形成建模。我们很难将云与圆或椭圆类比，但是它们确实很像分形。


曼德尔布罗特集

对于实用性问题，分形也能给我们带来新的希望，例如在地理制图中。你可能会认为海岸线长度的测量是一件简单明了的事，没什么特别之处。但实际上，你得到的答案将取决于测量时所用的单位。当你选择千米作单位时会得到一个答案，但是当你选择米作单位时就会得到一条更长的海岸线。这是因为较小的单位能够比较大的单位更好地与海岸线上的角落和缝隙贴合。如果我们用的量杆长度等于一个原子的直径，海岸线实际上将会非常长。在雪花曲线的每个生成步骤中，突起越来越短，但总长度越来越长。

我们的多维世界

从古希腊人只考虑一维、二维和三维世界时起到现在，数学家们已经揭示出有其他更多种维度（也包括分数维）活跃于其中的场景和概念。现在，“纯粹”数学家可以随心所欲地使用任意多个维度，重视实践的科学家们也感到需要从这个将三个维度强加于自身的世界中逃离出去。非但如此，到了21世纪，全世界各大学中的理论物理学家们在爱因斯坦的四维时-空连续统的基础上更进一步，致力于建立一种无所不包的物理理论。根据一种特定的宇宙观，宇宙由微小的“振动弦”构成，包含11个维度，这方面研究以极端奇异的方式呈现于弦论这一领域中，到目前为止，它看起来仍旧十分迷人，但却无法证明。无论是在数学还是数学以外的领域中，对于每一个热爱思考的头脑，维度概念始终保持着强大的吸引力。

13 蝴蝶的翅膀真能导致飓风吗？
——混沌理论，气象方程与奇异吸引子

尾蝶是蝴蝶中最美丽的一个物种，生活在人迹罕至的亚马逊河上游地区。它轻舞双翼，享受着巴西阳光的暖意。它不会知道，这个简单的动作会导致北方几千英里之外的一阵飓风。这个生动的画面成为"混沌理论"的一个鲜明的广告，激发出公众的想象力。它推动了一个新的研究领域的创建，这个新领域与相对论和量子力学一样，已经被推崇为20世纪科学和数学领域中最重要的进展之一。

让我们来做一个实验。在早上7点时，将一个机械钟调准时间，上好发条，让它开始运转，与此同时，在它旁边放一个调好时间为7:01的钟。很自然，在一整天里，第二个钟一直会比前一个领先一分钟。这就是我们所理解的机械论宇宙的样子，像钟表一样运转。但是并非所有的系统都以这样的方式运转。例如，同卵双胞胎的成长发展就远不是可以预测的。刚刚降生时，他们在外貌和行为表现方面可能都没有什么明显差别。在生命的前几年中，他们的生活可能会以相似的方式前行，但是，个体性格终究会逐渐成形，他们的人生轨迹也会因此而偏离。到成年时，他们可能已经过上了完全不同的生活。

从基本的层面来讲，我们已经对“混沌理论”做了一番浅明的概述，该理论主张，随着时间推移，系统发展遵循一条衍变式的路径，而初始条件的微小变化或差异都会使这条路走向不同的方向。从本质上讲，进化学说就是对我们所处世界的混沌理论的最强有力的诠释。不过从数学的角度来讲，混沌理论的普及主要还是发生在我们自己的时代，而不是达尔文的时代。

喝咖啡时间和蝴蝶

20世纪60年代，美国数学家、气象学家爱德华·洛伦兹正致力于研究可以用来预测天气的纳维尔-斯托克斯方程组。这些方程用三维空间中的点的x, y, z坐标来表示气象流的速度，因此求解这些方程就意味着寻找这些点所描绘出的轨迹。洛伦兹只能满足于用计算机求出数值解，因为要找到具有精确形式的解还完全没法做到。他将数值技术所要求的数以百万次的繁重计算任务交给计算机来做，在喝咖啡时间来到之前，他要先将x, y, z的初始数值输入计算机。回来时，他发现计算机已经算出一种方式的衍变路径——但是其中有一个问题。这一路径与他早先用同一组方程在类似实验中得到的结果不同。

> “当然，对于很多作家来说，可能更愿意用混沌这个词来表示一种模糊的感觉，但在物理学中，它指的是一种特殊的现象，即，在一个非线性系统中，输出结果常常是不确定的，极容易因为初始条件的哪怕是细微的变化而受到影响。”
>
> **穆雷·盖尔曼**
> **诺贝尔物理学奖得主（1969）**

经过一番思考，他恍然大悟，原来在第二次实验中，他将早先用的初始值x, y, z四舍五入了。比如说他原来使用的x值可能是5.890 453 3，但是后来他通过四舍五入将小数位数稍微减少了一点，变成5.890 45，然后才去享受喝咖啡时间。实际上，计算机根据两组略有差异的初始数据计算出两条衍变路径。

这一微小的数学差异经过洛伦兹的发挥改造，变成关于混沌理论的一个不朽的比喻，出现在他1972年的演讲“可预测性：一只蝴蝶在巴西舞动双翅，就能引发得克萨斯的龙卷风吗？”之中。他试图表达的一个关键思想就是*敏感性*。我们前面提到过，分别被设定为早上7:00和7:01的两个机械钟并不具有这种敏感性，因为它们的步伐节奏并不随时间推移而改变。（当然，我们知道在*现实世界*中，机械动力总会在某个时刻用完，在重新上发条之前，其中一个钟会比另一个持续得久一点，不过我们可以抛开那些问题不谈：在这里，问题的关键在于发条装置，以及它所遵循的机械原理。）但是在双胞胎的例子中，刚出生时的微小差异会随着时间流逝而放大。用专业术语完整地表达，“对初始条件的敏感性”意味着在时间零点出现的微小差异就会导致后来的巨大差异。在气象学领域中，洛伦兹注意到下面的事实：即使是在初始数值彼此间极为接近时，某些气象方程也会预测出完全不同的天气。

庞加莱和混沌的起源

实际上，在洛伦兹进行那个演讲之前，“蝴蝶效应”就已广为人知。早在六十多年前，昂利·庞加莱就已了解这种现象，甚至还设想过进行天气预测，他曾这样写道：初始条件的微小差异就有可能导致最终结果中的极大差异……前一步结果中的微小误差在后面的步骤中可能就会变成巨大的误差。不过对于庞加莱来说，没有计算机的运算能力作为辅助，这些“微小差异”恰好阻碍而不是辅助他

作出预言，他认为：预测是不可能的。

庞加莱是在研究天文学中的三体运动问题时注意到上述现象的。用数学方式描述协同运动的两个物体运动的问题（例如月球环绕地球运动）是比较容易的。庞加莱面临的挑战是对三个物体进行同样的研究，这个问题可以是月球、地球和太阳的动力学，也可以是亚原子水平上一个氦原子的动力学，因为它包含一个原子核与两个依轨道运转的电子。

在他所研究的动力学问题中，庞加莱实际上已经发现了“混沌”。随着时间推移，三体运动相对于之前的运动轨迹而略有偏移。他的研究揭示出一种新的数学现象，并且很快吸引了其他人的关注。在美国，乔治·伯克霍夫将对于动力系统的研究纳入拓扑学领域之中，拓扑学是富于纯数学味道的几何学领域，它关心的问题包括点与点之间的连接方式，以及一个表面能否变换成另一个等（参见第17章）。

> “为什么气象学家在以某种确定性来预测天气时会遇到如此大的困难？……哪怕是在任意一个点处出现只有十分之一度左右的偏差，也会使暴风在这里而不是另一个地方出现，它的破坏力就会延伸到几个也许本可以得以幸免的国家。”
>
> **昂利·庞加莱**

然而混沌并不是一个完全抽象的研究主题。尽管庞加莱是在天文学研究中与它相遇的，但是混沌在各种实用学科中也一样会展示自己的影响力。一个可以用方程来描述的电二极管也同样呈现出对于初始值的敏感性。一个基于混沌理论的新型研究领域出现了，它将诸如力学、无线电振荡、控制论与纯数学等表面上毫无关联的不

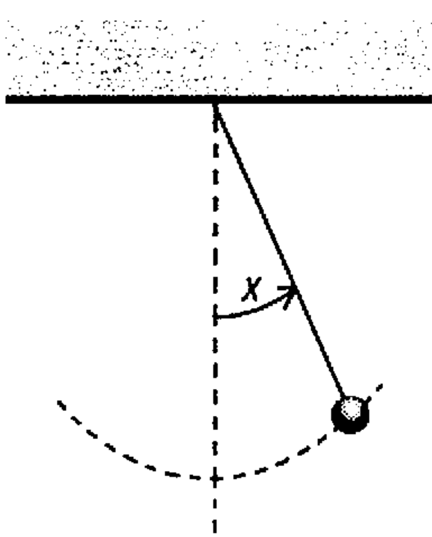
同学科联系起来。

奇异吸引子

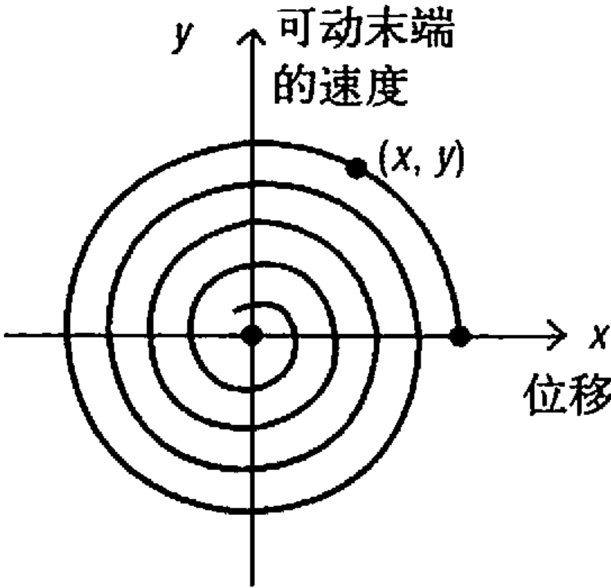
庞加莱用抽象点空间中的一种新型图示方法来描述一个动力系统的不同的运动相。在他看来，这种相空间是了解动力系统的一扇重要的窗口，现在让我们通过常见的钟摆来看看这种方法是怎样起作用的。

单摆是一根刚性杆，它的一端连接在一个固定点上，另一端是可以摆动的，整根杆可以来回摆动。描述单摆运动的变量是x，杆与竖直方向所成的夹角，和y，即摆动的角速度。因此，不妨设开始时y的值是0。如果我们将单摆从某个位置S开始释放，那么单摆将逐渐损失能量并最终停下来，因此从长远的角度来看总会有$x=0$和$y=0$。它的运动可以用二维平面中动点(x, y)的轨迹来描述。无论开始时可动末端的初始起点S在哪里，单摆都会越来越慢，所能达到的高度也会逐渐下降，最后螺旋式下降到原点$(0, 0)$，这些全都是可以预测的。

在这个例子中，“吸引子”就是最终停止的位置——二维平面中的原点$(0, 0)$，这一点处的位移和速度也都是零。对于摆的任意起始位置，这个点都会吸引描述单摆运动的螺旋线。



单摆



单摆的二维相空间

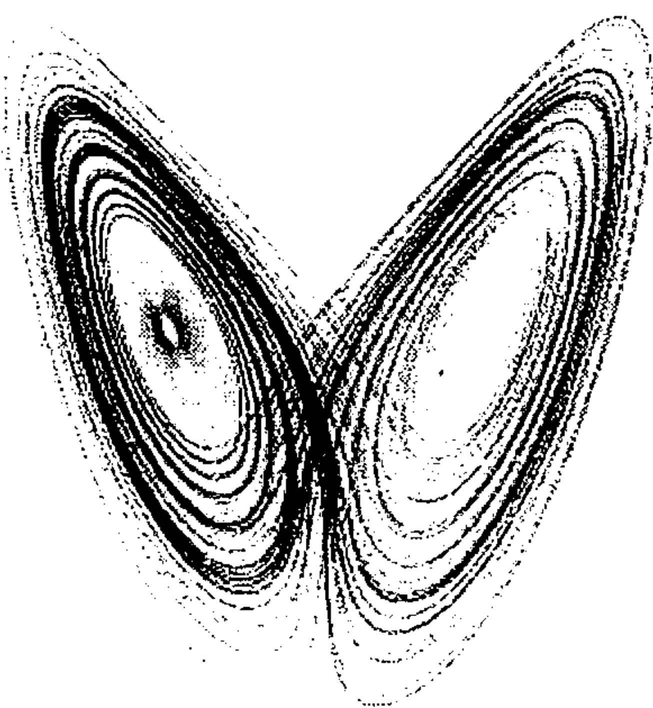
古老的落地式座钟里面的摆略有些不同，因为它会一直不停地来回摆动。描述其运动的轨迹不是终止于原点的螺旋线，而是沿圆

周运动。如果它的运动受到些微扰动（比如说座钟被撞了一下），那么描述单摆运动的轨迹就会被吸引而回到圆形。

上面讲的这两种类型都不符合“奇异”的数学意义。这个形容词的意思是说，吸引子既不是一个点也不是一个圆，用该领域的先行者大卫·鲁尔的话来说，是“一个奇怪的点集”，它们散布于空间之中并且吸引运动轨迹。用显然不那么数学味儿的语言，鲁尔称它们是“由点组成的云”，有时也会称之为“星系或烟火”。在他看来，“奇异吸引子”这个词“非常适用于那些我们对其所知甚少、令人惊讶的对象”。这个名字流行起来，并且很快就激发了科学界和一般公众的想象力。

在洛伦兹研究气象方程时所遇到的例子中，由他的气象方程所决定的解轨迹就围绕在这样一组分布在三维空间中的“奇怪的”点附近。由此而得到的可视图象（洛伦兹吸引子）成为新的混沌科学的标志性图象。

但是还有个问题没有解决：洛伦兹吸引子真的*存在*吗？它究竟是数值方法所产生的人为产物，还是洛伦兹方程本身的内在属性？在洛伦兹的发现出现30年后，这个问题得到了回答。答案是肯定的：洛伦兹奇异吸引子存在，而且它是一个分形（参见第12章）。一旦确定了这一点，下一步就是测量它的分数维度，结果大约为2.05。



洛伦兹吸引子

种群与混沌原理

混沌所关注的是一个时间段中的发展。在洛伦兹研究的气象方程中，时间是连续流动的。不过，在更简单些的方程中

也可以找到混沌理论的身影，其中时间是用离散值0, 1, 2, 3…来度量的，例如0年、1年等。种群增长的理论长期以来一直是用方程来描述的，但是此前还没有人从混沌的角度来看待这个问题。

在最简单的种群模型中，某一给定年头开始时种群的数目是前一年种群数目的若干倍。如果这个倍数，或者称为增长率（比如说）是$r=2$，并且在某一年开始时种群数目是50万，那么在下一年开始时种群数目将是它的二倍，也就是100万；而再下一年开始时的种群数目将是200万。利用这种逐步进行的机制，我们按照$n=0, 1, 2, 3\cdots$来计数时间。事实上，我们总可以用数学公式来表达，利用下列基本方程：

$$P_{n+1}=r\times P_n$$

根据前一年的种群数目P_n来求得下一年的种群数目P_{n+1}，每步推算应用一次。

但是，这种类型的方程能够产生一个很不现实的结果。由于种群是按照比例增长的（这决定了方程是*线性的*），因此该种群会一直持续增长。人口普查结果和历史事实告诉我们，这根本是不现实的。我们需要更好的方程。

一种改进结果是方程

$$P_{n+1}=(r-sP_n)\times P_n$$

其中，新的增长速率$(r-sP_n)$依赖于当时的种群规模，因而会随着种群增长而减小。假设和刚才一样，开始时种群数目是50万，$r=2$，我们选取（比如说）$s=0.04$，根据上面的方程我们可以计算出，下一年开始时的种群数目为990 000，再下一年则是1 940 706。可以看到，这两个结果比基本模型所预言的值（100万和200万）要小。

有了这个方程，我们很容易得到一个改良的版本，即（大名鼎鼎的）逻辑斯蒂形式：

$$x_{n+1} = r(1 - x_n)x_n$$

其中，变量x_n的取值范围是从0到1，但是如果需要的话，我们很容易逆推出实际种群规模，方法是将我们的答案乘以r与s的比值，比如这里是$\frac{2}{0.04} = 50$。不过现在我们更有兴趣的是方程本身，它将我们带入了混沌的领域。

x_n与$1 - x_n$的乘积决定了这个方程是非线性的，这也正是我们利用这个方程来讨论混沌概念的原因。从数学的角度来讲，混沌是非线性方程的性质。

如果还像刚才那样令$r = 2$，无论第零年时的初始值即x_0是多少，年复一年地计算下来，种群数目都会在较长时间后被吸引向一个固定的值0.5。（例如当$x_0 = 0.6$时，逐步迭代的结果是$x_0 = 0.6$，$x_1 = 0.480$，$x_2 = 0.499$等，并且快速归结为0.5这个值。）用混沌的语言来讲，存在唯一一个数值0.5将所有的数值序列“吸引”到自身：它就是吸引子。

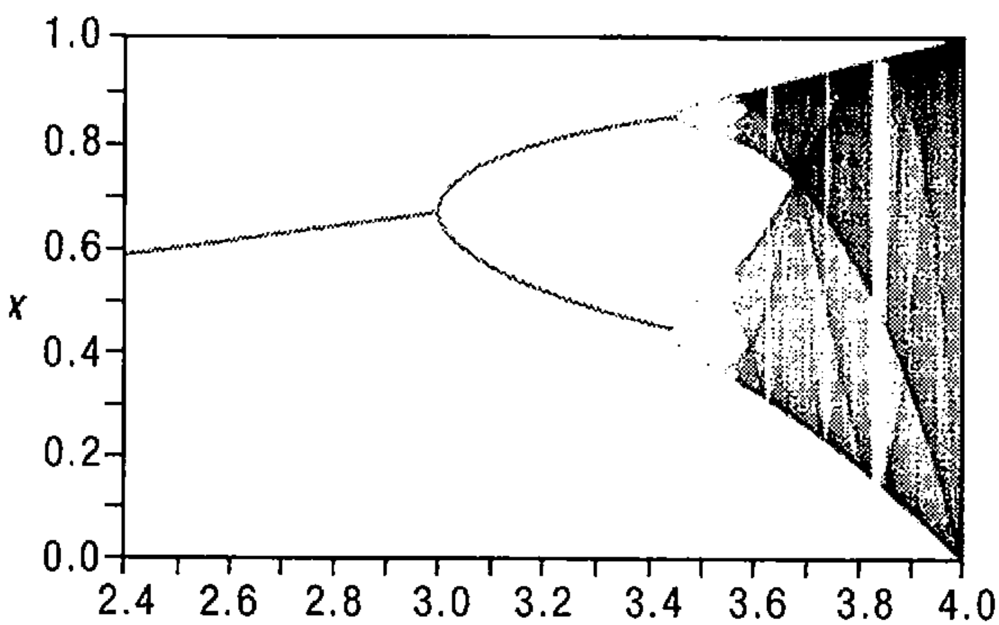
如果我们给逻辑斯蒂方程中的r带入一个略大于3的值，就会发现长期计算后可能得到的数值不是一个而是两个，于是，种群数量将在它们之间年复一年地反复振荡。这是因为这个r值所对应的（所谓的！）“不动点”具有不稳定性。例如在$r = 3.2$时，两个吸引点近似为0.513和0.799，于是（还是假设基值$x_0 = 0.6$）逻辑斯蒂方程将会得到相继各年时的值0.768，0.570，0.784，0.541，0.795等，游动于两个吸引点之间。

通过进一步增大r的值，我们发现吸引点的个数不断翻倍，于是得到4，8，16和32，继续增大r的值，就能得到更多的吸引子。不过，

随着吸引子的数目越来越大，预测也就变得越来越难。当增大到$r=3.57$甚至更大，一直到4时，就不再有任何可识别的模式，因而不可能进行预测。调节r的数值类似于开启水龙头。开始时水滴是规则且可辨识的，但是随着龙头逐渐开启，水滴逐渐变成无法分辨的水流。

对于较大的r值，蝴蝶效应才会真正出现，这时，即使是极为接近的初始种群数量也会随着时间的推移而以截然不同的方式发展。例如，我们令$r=4$，并将其应用于非常接近但又不是完全相同的两个初始种群数值$x_0=0.600$和$x_0=0.610$，可以看到，才经过最初的几个结果，刚到第六年（x_6）时，它们各自所产生的种群数值就已经是0.025和0.529。

“混沌”的发散模式可以用“分枝图”表示出来。分枝图显示出分散的吸引点是怎样随着r值增加而产生出来的。将起始值x_0标记在竖直轴上（还是取$x_0=0.6$），分枝图显示出不动点如何在不同r值处分裂成叉子形的两个分枝。于是当r达到3时，只存在着唯一一个极限值，但超过3以后，吸引子的数目就会增加到2，4，8…直至进入混沌！



呈现混沌发散的分枝图

混沌的爆发

自混沌概念浮出水面以来，见诸于世的研究著述可谓汗牛充栋，它也由此而成长为一种跨度广阔、内容丰富的理论，而且现在仍在快速发展中。它在数学建模中的应用也是如鱼得水。今天，工程师们通过详细考察支配船只在深海中运动的非线性方程，可以更好地

模拟船只的摇摆晃动，以此来预防船只颠覆的危险。在从金融市场的全局活动到心搏节律等一系列问题上，混沌理论和非线性方程都打上了自己的烙印。

尽管上述应用给人留下了如此深刻的印象，基于对混沌之理解而产生的创造力又是如此强大，限制总还是存在的。混沌原理依赖于对初始条件的敏感性，这意味着要预言某些物理过程在理论上是不可能的。长期天气预报就属于这类问题，因为这种预报需要及时对某些位置的天气条件有精确的了解，而这是我们不可能确定的。这种不可避免的不准确性足以导致预报出错。

也就是说，这完全取决于巴西的燕尾蝶是选择要还是不要扇动它的翅膀。

14 我们能创造一种不可破解的密码吗？——密码术，密码机与量子计算机

1843年，当身为艺术家和企业家的塞缪尔·莫尔斯将他的第一句编码信息“上帝的成就如此辉煌”从华盛顿发往巴尔的摩时，一种出色的密码诞生了。它将在全世界得到广泛应用，即使是在那些没有电报可用的地方。当然，作为一种编码，它被设计成一种简明易懂且非常实用的传输信息的方法。不过，在历史上还存在着与它完全不同的一片天地（保守秘密的编码方法和密码术）使我们沉迷于其中。从古代战争到21世纪的计算机系统，数在编码和解码活动中都起着重要的作用。

尽管说通信在人类活动中处于核心地位可谓老生常谈，但是作为社会化的动物，我们的确很难想象一个没有通信联系的世界。日常语言本身就是一组编码——使用习惯和规则条例决定了谁能理解它们，谁不能理解它们。不过在使用“编码”这个词时，我们通常还会想到别的含义，编码应该更富于目的性，因而可以与日常语言区分开来，在很多情况下它应该是更有针对性，更细致，甚至是更隐秘。从那种意义上来讲，编码成为了特定群体与个人（在计算机时代指的是机器）之间通信的方式，而窥视的眼睛则是不受欢迎的。

从密码术到编码

一种曾在历史上流行的编码方法就使用了一种密码术。这种方法是将信息中的字母根据某种系统进行变换，或者是将字母替换为另外一组符号。该信息（或称“明文”）即被加密而成为“密文”，接着再由接受信息的人来解密。

一个早期密码的实例，是尤里乌斯·凯撒曾使用过的“移三系统”。在将信息加密时，把明文中的每个字母都用字母表中位于其后第三个位置上的字母来替换。于是，字母*a*就用*d*来代替，*b*就用*e*来代替，依次类推。当我们替换到*x*时，字母表用完了，这时用*a*来替换，就像将字母表重新开始一样。例如，将明文“enemy in next town”加密，就会得到“*hqhpblqqhawwrzq*”（没有考虑空格）。

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
d	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*	*a*	*b*	*c*

移三密码系统

随着这种质朴的“凯撒移位”密码逐渐广为人知并且因此而变得越来越不安全，人们需要发明出新的也更为复杂的加密方法。在更安全的密码中，人们不再只是单纯移动字母表，而是用更巧妙的方式来调度字母。

当加密过程涉及将整个词或词组中的字母替换时，我们使用专门术语，称之为“编码”。与密码术不同的是，编码需要一本关于密语的“字典”，无论是发送者还是接受者都要有一本字典。在一般情况下，我们经常会混用“编码”和“密码”。

编码，密钥和一位希腊历史学家

随着编制密码的高级方法逐渐普及，专业密码破译者的出现也

就成为必然。通常的密码破译者应该是数学家、统计学家、语言学家和解谜者的结合体。他需要富于想象力、有条不紊、锲而不舍，还要有挖掘蛛丝马迹并且利用它们来解译信息的能耐。破译者必须熟知加密者的各种技巧，还应随时了解编码学的新发展。

破译者知道，一种有效的编码必须要有规则，而潜在的弱点也就隐藏于其中——如果能发现规则的话，就有解码的可能性。如果将字母用其他字母随机地进行替换，就可以对一条信息进行加密，但是未来接受该信息的人却没法解密：那根本就是浪费时间。常用的做法是将规则转化为一把“密钥”——一种掌控全局的原理或设计方法，以开启密码。因此，有必要通过另外一些在历史上很重要的密码来了解一下密钥的作用方式。

在第一次世界大战中，密码制作者喜欢使用“波利比乌斯的5×5方阵”，这个名字来自于希腊历史学家波利比乌斯，他曾以这种方式用一对数字来将信件加密。在这个方阵中，每个字母对应于一对数字，其中一个来自于横行，一个来自于竖列。例如，E处于第1行第5列，因此E就对应于15。利用波利比乌斯方阵，就可以将“enemy”这个词加密为1533153254。很显然，任何人只要拿到这个方阵就可以破译信息，因此它也成为这一系统中的密钥。

	1	2	3	4	5
1	A	B	C	D	E
2	F	G	H	I/J	K
3	L	M	N	O	P
4	Q	R	S	T	U
5	V	W	X	Y	Z

波利比乌斯的5×5方阵

1918年，弗里茨·尼贝尔上校设计出一种方法，他将波利比乌斯方阵中的数字1, 2, 3, 4, 5改写成字母A, D, F, G, X。这几个特殊的字母具有不同的莫尔斯电码信号，因此在传输时不大可能与其他信

号混淆。利用这种经过“安全性升级”的波利比乌斯方阵，我们将字母表随机地排布于方阵之中。

	A	*D*	*F*	*G*	*X*
A	B	X	W	K	C
D	O	Z	A	I/J	Y
F	T	F	U	P	M
G	N	L	Q	H	R
X	S	E	V	D	G

弗里茨·尼贝尔对波利比乌斯方阵的改进

于是，信息“enemy in next town”就会经过加密而变成令人困惑的“*XDGAXDFXDXDGGAGAXDADFAFADAAFGA*”。作为进一步修饰，在发送之前，还可以引入二级密钥将加密信息打乱。但是，具有这种类型密钥的密码却包含着固有的弱点，因此，几乎在举手之间即可将其破解。

机器与数学家

利用诸如波利比乌斯方阵等方法用人力将信息加密是一件很花时间的事情，而且还可能引入错误，尤其是在战争或者战役等紧急情况中。它本身是一个机械过程，因此干嘛不直接用机器来完成呢？“伊尼格玛”（“Enigma”，来自于希腊词汇“谜题”）是一种德国机器，于1923年问世，它最初投入市场时并不是用于军事目的，而是为了传输商业敏感材料。在第二次世界大战中，它成为布里奇利花园中的密码破解者取得成功的象征，这些密码学家经过努力将它的秘密揭示了出来。

“伊尼格玛”通过机械器件与电子电路的联合装置（它是一种电子—机械装置）来输入。其核心部分是排列在同一根轴上的三个圆盘（转轮），每个转轮上标记着字母表上的26个字母。在每个圆盘的一面上的电接触都与另一面上的电接触用导线连接，每个转轮上连着不同的导线。

机器上有一个写有字母的键盘区，还有一套字母符号，每个都

能发光。将A键按下时，一个不同的字母（比如说P）变亮了；于是第一个转轮转动一步。按动26次之后，第一个转轮转动一周，于是第二个转轮转动一步；按动26次之后，第三个转轮开始运转。每按一次键，机器内部就产生一个不同的电路。如果发信人以后必须要输入字母A，还可以保证产生不同的编码方式，因此并不是必须要点亮P。正是这种“多字母”替换方式赋予了“伊尼格玛”相当的复杂性以及与密码破译者对抗的力量。通过计算，对于基本模型，要经过26×26×26＝17 576次按键之后，电路才会重复：一个令人望而却步的可能性阵列。

在理想情况下，一台加密机器不应该呈现出任何使破译者有机可乘的特征。这正是“伊尼格玛”的弱点所在。即使不考虑由转轮所导致的周期性质，它还是有两个缺陷。

(1) 无法将一个字母加密后变为自身，因此A加密后得到的一定是不为A的字母。

(2) 伊尼格玛是可逆的。如果A加密后变为P，那么（对于同一位置的圆盘）按下P就得到A。这是有利于信息接受者的，但也同样有助于破译者。

发信者给机器中每个圆盘上最上方的三个字母设置一个初始状态——例如，可能是RSM。我们假设发信者要发送的明文信息还是“Enemy in next town”，而接受者得到打乱后的密文PBNTAMJGLQWDLFZ。要将其解密，接受者同样将“伊尼格玛”设定为RSM并令其按相反顺序工作，输入密文以复原最初的明文。很明显，一件重要的事情是：接受者必须知道发送者的初始位置，因此预先有准备的接受者手中要有一个密码本，其中记录着每天要用的密钥。

如果认为17 576个起始位置还不足以提供足够的安全性，为防止被解译，还可以快速增加可能性——而且是增多得令人震惊。这正是数学得以插足其中的原因。利用三个*固定不动*的转轮可以增加初始位置，可以选择6＝3×2×1种可能的在轴上放置转轮的方式。于是，增多后达到6×17 576＝105 456种有效的起始位置。如果是从五个转轮中选取三个的话，则可能的起始位置会进一步增加到10×105 456＝1 054 560种。

所有这一切都还没有算上安装在“伊尼格玛”前面的线路连接板。这种方法是通过*互换*在线路板上的引线而将字母表中的某些字母成对交换。对于每一起始位置，只从字母表中选取三对字母交换就会产生3 453 450种布线方式，而选取十对字母进行交换，则会对每一起始位置产生令人惊讶的超过1 500 000亿种布线方式！

考虑到可以用来加密的方式的规模，对于密码破译者来说，关键在于找到一个“对照”，那是一个入手的词，它可以提供关于密钥的线索，使我们得以排除那些错误的尝试。例如，消息可能总会以“最高机密”或“嗨希特勒”为开头。

德国最高指挥部成员之间的通信是用洛伦兹机来加密的，这是一种与电传打字装置联合使用的加密机。它包含至少12个可用的转轮。要将这些信息解密，用到了世界上第一台编程计算机“巨人”，它是由英国电信工程师托米·弗劳尔斯设计出来的。

公钥加密

如果说是密钥（无论是机械方式还是其他方式的）的存在性导致了编码与密码在安全性能方面的弱点，那么解决方法是什么？当答案出现时，令人大吃一惊——将密钥公开！当然，也不是完全公布，还是有某些重要信息是只有接受者知道的。

如同大家都知道的那样，“公钥加密”首先要分配选定的数。我们可以通过一个实例来了解其原理，比如说，我们能够使用的所有数就是3和55。它们可以在收音机中广播或者是在地方图书馆中查到。让我们假设早已熟悉的信息“enemy in next town”在一本同样是随处可得的密码本中被编码为“18”。如果一位窃听者知道该信息为“18”，只需查阅密码本即可得知其内容。

但是，利用3和55这两个数我们可以将信息“18”先加密再传输，即可对偷听者保密。这时一些非常古老的数论知识就有了发挥作用的机会，这些知识曾被人认为是不可能具有任何实际用途的。利用3这个数，我们先计算：

$$18^3 = 18 \times 18 \times 18 = 5832$$

现在我们将这个数除以55，并计算所得的余数。因为$5832 = 106 \times 55 + 2$，所以余数是2。现在我们将2作为“18”的编码结果并将它进行传输，要记住的是，任何人都能对信息进行编码，因为3和55这两个数是公开的。

于是，现在接收者读到了2，并且要将它解码。关于55这个数，接收者知道一些并非人尽皆知的事情——它是两个质数的乘积，即$p = 5$和$q = 11$。这时该由费马所发现的定理和中国剩余定理联手来发挥作用了，这是纯粹数论魔法的有力结合。利用另一个公开的数3，定理保证会存在具有如下性质的一个数x：$3 \times x$除以$(p-1) \times (q-1)$时余数为1。于是，接收者要对于两个质数分别是$p = 5$和$q = 11$，即$(5-1) \times (11-1) = 4 \times 10 = 40$的情况，计算出这个数$x$。

我们知道这样的数是存在的，但我们还必须把它计算出来。在这个例子中，用较小的数进行计算，我们看到$3 \times 27 = 81$，而81除以40的余数是1，这意味着x的值就是27。这就是我们在将接受到的数2

解码时所需要的数。

> “就算用全世界所有个人电脑（大概2.6亿台）一起来破解用P[retty]G[ood]P[rivacy]加密的信息，那么平均起来每破解一条信息，还是需要大约为宇宙寿命的12 000 000倍那么长的时间。”
>
> **威廉·克罗威尔**
> **美国国家安全局副局长，1997年时对于PGP公钥加密系统的描述。**

为此，我们先要计算$2^{27}=134\,217\,728$。将它除以55，得到2 440 322，余数为18，因为：

$$134\,217\,728=2\,440\,322\times 55+18$$

余数就是原来的信息“18”，至此解码成功。

公钥加密系统的安全性取决于接受者所了解的保密信息，在我们这个例子中就是$55=5\times 11$。对于像55这样很小的数，要找出质因数5和11是很容易的。但是对于一个有1000位数字的数，就像实际应用中所选择的那样，找出它的质因数几乎是不可能的。将两个质数乘起来要比寻找给定数的质因数的相反运算容易得多。对于任意给定的数，利用有效的计算机算法来回答“这个数是质数吗？”这样的问题是可能的。但是找出它的质因数则是完全不同的问题，而且至今也没有有效的计算机算法能够实现这一目标（参见第3章）。一台家用计算机可以将一个20位的数进行质因数分解，但是要对一个100位的数进行同样的运算可是一场艰苦的战斗。随着计算机的能力越来越强大，公钥加密系统可以选取超越现有分解能力之外的更大的数。但是我们可以始终做到这一点吗？

在1978年建立的McEliece密码系统中，发布出来的公钥不是单独一个或者一对数而是一大批数。在实际应用时，建议这一批数（称

为一个“矩阵”）应包括664行及1024列。这个矩阵用来加密信息，而接收者还了解一些额外的信息，即如何将它拆分成三个子块，这个拆分对应于将数分解为质因数的过程。尽管McEliece密码系统在实际应用中不太容易操作，但是它有一个可取之处：现存方法对它几乎无计可施。

量子解码机

看起来，公钥加密已经是加密技术发展的极致了。但是，我们还能继续前行多远呢？我们还需要获得更强的保密性吗？量子计算机是一种新兴技术，它的计算能力有可能会使公钥加密系统中的质数方法失效。尽管“古典”计算机没有能力找出极大数的质因数，但是目前处于设计阶段的量子计算机却能提供更高效的计算能力，因而有可能将质因数找出来。如果真能实现的话，问题就严重了，因为这意味着目前用来维护电子邮件和网页安全的编码系统是可以破解的。

那么，我们能创造一种不可破解的密码吗？目前的回答是“能”，实际上我们现在就有一种：McEliece密码系统目前仍是安全的，即使是尚处于理论阶段的量子计算机也无法破解它。不过也必须承认，还有另一种回答是“走着瞧吧”。在编码者与解码者之间的战斗中，赌注一直在持续增加。

15 数学美吗？
——音乐、艺术、黄金数与斐波那契数列

谈到数学中美的概念，可能有人会感到好奇。我们很容易感受到莱昂纳多·达芬奇的画作与莫扎特的曲作品中的美。但是一个艰深的定理美在哪里？更不用说还有那满纸的陌生符号了。数学可以带来意料之外的惊喜、面对困难的挑战、理解后的满足感以及公式结果的简洁明快。但是美从何来？尽管数学与美无疑是两件事，但是数学家们却始终保持着对于美的敏锐感觉，而艺术的历史也揭示出美的观念与数学原理之间的大量联系。

对于一个正在学校里与乘法表和百分数奋战的孩子来说，“数学是美丽的”大概不会有什么意义。而当一个数学家因为某个数学构造或定理之美而陷入狂喜之时，我们也似乎要在头脑中实现某种极端的转换之后才能与他产生共鸣。就其所包含的多种并存含义的所有层面而言，“美”这个词真的能够用于数学吗？

关于这个问题，让我们先来定义美吧。现代人往往可能会认为它是主观的。哲学家大卫·休谟告诫我们“事物中的美只存在于思考它们的头脑之中”，这是柏拉图的观点“美存在于观看者的眼中”的一种变体。但是美学家和思想家并不总是同意这一观点。对于伊曼纽尔·康德来说，美是一种客观属性并且普遍存在，对于诸如时装模特和饰演男主角的艺人来说，看起来在公众生活中无疑存在着一种约定俗成的“美”的概念。一位19世纪的数学家，即研究出了

八元数的亚瑟·凯莱（参见第5章），把柏拉图当做自己的指路明灯，但他不打算参与这场特殊的纷争，称："就这一点而言，一条数学定理与任何其他事物是一样的：美是可以感受但不可以解释的。"

尽管难于言说分明，但是对于科学家和数学家来说，对美的敏感一直是一种强烈的激发因素。20世纪最伟大的数学物理学家之一保罗·狄拉克有一句名言是这样说的："一个方程中包含着美要比能够与实验吻合更为重要。"也许这种看法有些极端，不过狄拉克也确实有资格说这句话，毕竟，他曾提出的那个方程是量子力学发展历史中最重要的事件之一。

但是，我们也不是浑浑噩噩地漂荡在一个无法估量的世界之中。如果我们尝试着深入到艺术的表象之下，从音乐到建筑再到绘画，就会看到那些经常被认为是对于美而言至关重要的原理，它们在本质上却都是数学原理——也许其中最具代表性的就要数对称和比例了。

音乐与数学

数学与音乐经常被看做是彼此关联的努力和才能——实际上，这两方面的技能已经被归结于大脑中的同一区域。古希腊时，毕达哥拉斯学派的学者就已经了解关于声音和振动弦的基本物理知识，他们指出了音阶与整数比例之间的联系。例如当最高音与最低音满足比例$\frac{2}{1}$时，就构成了一个八度C, D, E, F, G, A, B, C，较高的C调是较低八度中C调的二倍。在两者之间，C大三和弦中音符的音高——C, E, G，与C的比例是$\frac{5}{4}$和$\frac{3}{2}$。以赫兹为单位进行测量，若较低C调的音高是261赫兹，则G调的音高是它的$\frac{3}{2}$倍，即391.5赫兹。

通过调节比例，音乐可以产生和弦与不协和，前者对于我们从音乐中获取的快感而言具有特别重要的核心意义。在研究能够决定音乐接收效果的组织方式时，学者们还发现了别的数学原理，对称性（毋庸置疑，这是一个数学原理）闪亮登场。在音乐中，力量和感染力经常体现于重复模式之中：相比之下，只出现一次听起来会显得很无力。

在编钟音乐中，可以清楚地看到数学与乐曲联系起来。其中，乐曲的属性是靠排成特定顺序的鸣钟声来决定的。“变化”的数学模式与近世代数中的置换群理论颇有相通之处。

美和图形

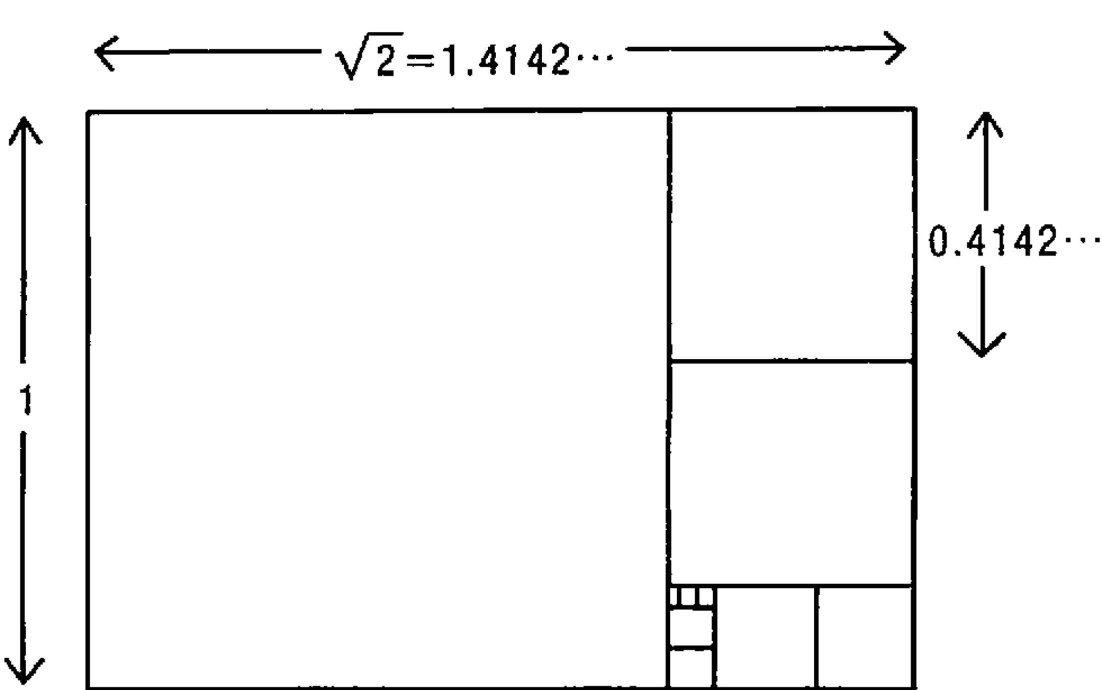
在谈论关于音乐的问题时，我们谈到了对称和比例。这些因素也是数学中关于形状与结构研究的重要组成部分。而它们最具魅力之处往往就在于其简单性，下面这个关于一个很简单的矩形及其性质的例子，就很好地体现了这一点。

> “我们在音乐中所获得的快感来源于计数，不过是无意识的计数。音乐，只不过是无意识的算术。”
>
> 高特弗雷德·莱布尼兹

最美丽的矩形是什么形状的？我们可能会说是正方形，即各边长度都相同的矩形。这是有道理的，因为正方形有很好的对称性，而且它还具有某些特殊的数学性质，例如两条对角线相交成直角。

但是还有其他一些研究矩形的方法，这时我们所关心的是它的比例。如果我们将一张矩形的纸片对折，并且沿着折痕展开，就会得到两张矩形的纸片。

如果较长边与较短边满足$\sqrt{2}=1.4142\cdots$比1的比例，那么较小矩形的两边恰好也会具有相同的比例。这是英国用来量度纸张大小的“A”方法的基础。任意一张A系列的纸张都是一个“根号2矩形”，原因就是它们与2的平方根的这种联系。



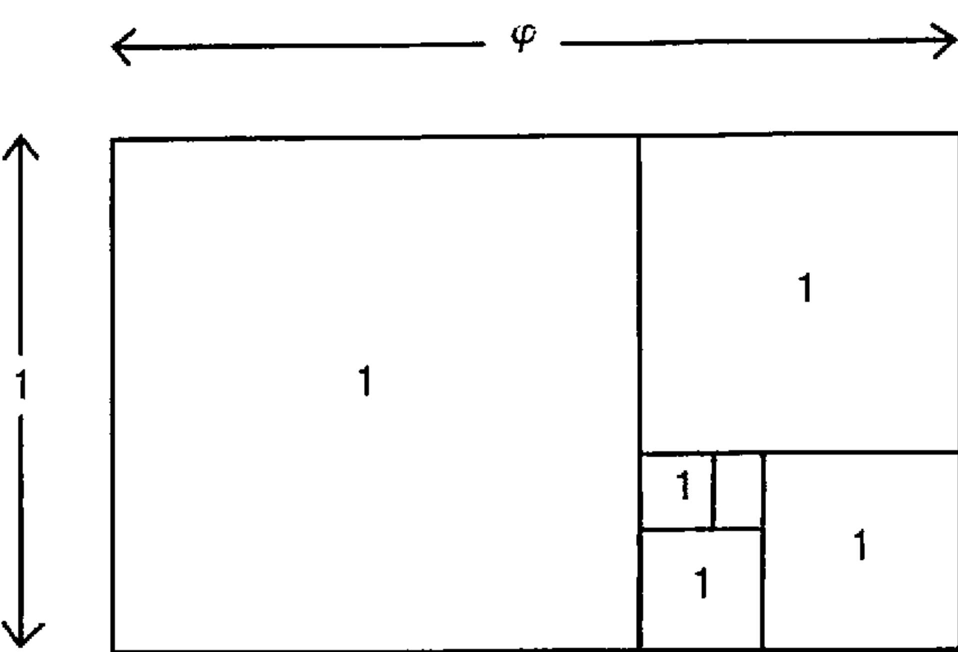
从一个根号2矩形中“取走正方形”

通过“取走正方形”可以对根号2矩形进行进一步的分析。如果我们从一个实际边长为1和$\sqrt{2}=1.4142\cdots$的矩形中取走一个正方形，就会余下一个短边为0.4142…而长边为1的矩形。

从这个矩形中我们可以再减去两个边长为0.4142…的正方形，接着就会余下一个短边为0.1716…长边为0.4142…的矩形。我们还可以从这个结果中再取走两个正方形，而这个去掉两个正方形的过程是可以无限重复的。从数学角度来看，该过程的另一种表示方法，即[1；2, 2, 2, 2, 2, …]，就是$\sqrt{2}$的连分数展开形式。于是我们看到，根号2矩形具有一种令人满足的对称性，但是它还不是非常完美。

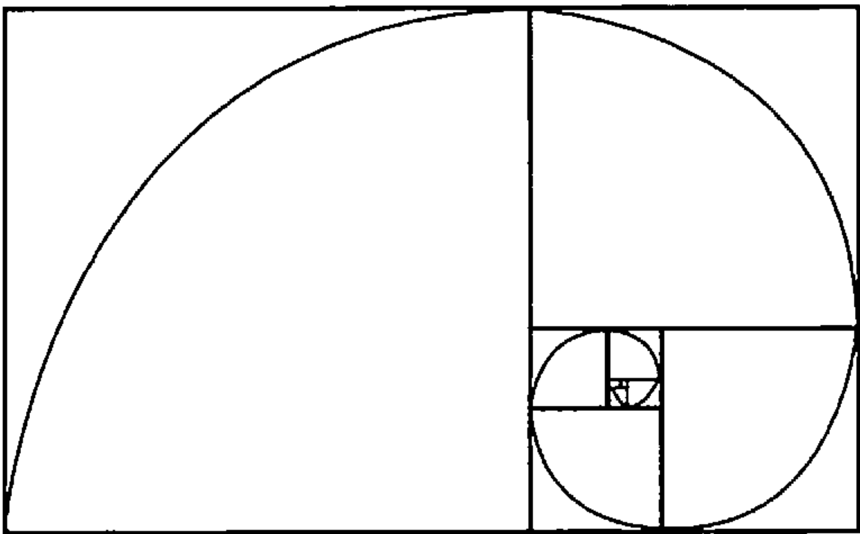
黄金矩形与黄金比例

在所有矩形中，处于最高地位的当属“黄金矩形”。对于一个黄金矩形，它的长宽之比不是$\sqrt{2}=1.4142\cdots$而是φ（phi），它的值是：$\varphi=\dfrac{1+\sqrt{5}}{2}=1.618\,033\cdots$数$\varphi$称为黄金数，或称黄金比例，它在数学中可以说是无处不在。



黄金矩形和其中的正方形

在一个黄金矩形中，如果我们减去一个正方形，那么余下的矩形还是一个黄金矩形。由于我们在每一步都取走一个正方形，因此φ的连分数展开形式是更特殊的[1；1, 1, 1, 1, 1, …]。



对数螺线

将黄金矩形中各正方形的会合点用一条曲线连接起来，产生了一个优美的图形。它接近于“对数螺线”，这使得该矩形的几何性质凸显出来，而它本身也是美丽的。

对于数学家来说，φ中存在着一种内在的美，它不需要依赖于详细阐述或进一步的应用。它的美是不言自明的。而在数学领域之外，黄金数φ所代表的比例关系对于西方文化来说是根基性的。由于它那无处不在的性质，φ曾被称为“神圣比例”，这个名字来自于卢卡·帕乔利的《神圣比例》(1509）一书中，他用测量人体的方式来描述φ，例如人的身高与从头顶到指尖距离的比例。

在艺术中，莱昂纳多·达芬奇将φ视为出现在自然界中的比例而为它痴迷不已。而在几个世纪之后，黄金分割派艺术家们抛弃了立体主义的正统观念，捍卫其神秘主义特征与经常富有象征意义的条理性。目前已经从黄金数的视角对几位艺术家进行了分析。超现实主义者萨尔瓦多·达利的名作《最后晚餐的圣礼》中就包含一个黄金矩形，因此达利在绘画中使用了数学图形。而具有讽刺意味的

是，杰出的“矩形的艺术家”，皮特·蒙德里安是一位出于某种原因而尽量避免受到φ影响的画家，他的作品中没有任何特别符合该比例之处。

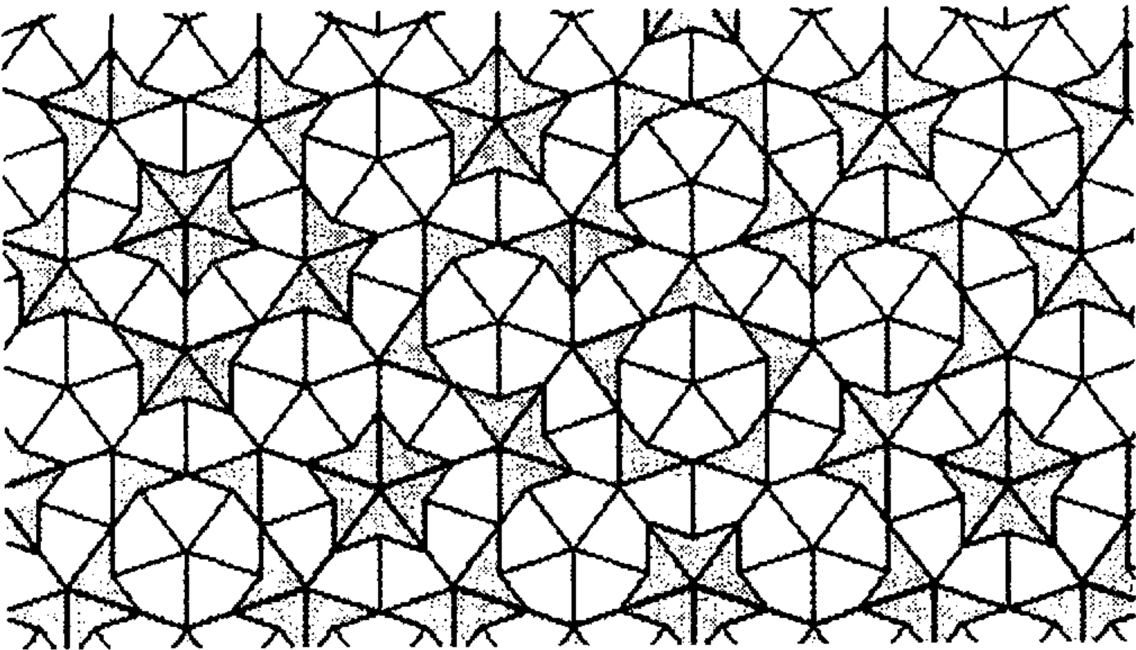
不过，黄金比例最能体现其无处不在而又历久不衰的影响力的领域还得说是建筑学，例如巴特农神庙中的比例。在这里，它有时也被称为“黄金分割”或“黄金中值”（golden section或golden mean）。到了20世纪，瑞士建筑师勒考尔布才明确地赞美数学比例对于他的现代派设计的意义，他的“模”理论是根据由黄金矩形构成的平面图来完成设计的。

镶嵌与几何学

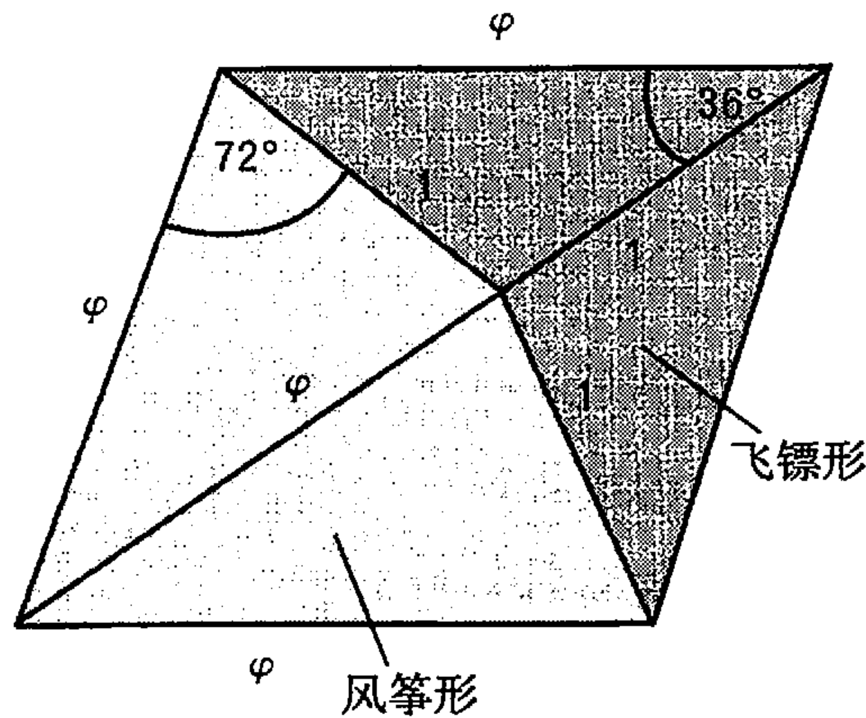
关于数学在艺术中的呈现，一个公开的范例是一类被称为*镶嵌*的有规律的装饰性图案：在一个平面内排布图形或瓷砖，使它们能够精确地合拢。在这类设计问题中，对称性扮演着重要的角色。

伊斯兰传统中的文化因素影响了镶嵌作品。伊斯兰传统通常反对在艺术或建筑中呈现先知穆罕默德或其他宗教界名人。其结果是他们反而拥有鲜明和丰富的几何图形传统。在西班牙，在14世纪摩尔人占领时期建成的格拉纳达的爱尔汗布拉宫中，也可以找到令人印象深刻的镶嵌作品。后人认为，这些特殊的实例影响了欧洲镶嵌艺术“大师”埃舍尔，他曾于20世纪20年代访问过西班牙。

如果一个镶嵌，比如说用六边形构造出来的镶嵌，不断重复自身，则称之为*周期性镶嵌*。这类镶嵌具有平移对称性，我们可以将其在平面上滑动，所得图案恰好位于原先的图案上。非周期性镶嵌不具有平移对称性，其图案不会重复自身，不过存在着某些仍具有“旋转”对称性的非周期性镶嵌（参见第18章）。


非周期性彭罗斯镶嵌的复杂实例

有好几种非周期性镶嵌是由罗杰·彭罗斯爵士发现的，其中有一种只用到两种图形，即风筝形和飞镖形，而它们又都是来自于菱形——一种具有两对互相平行的边且四条边长度都相等的图形。而风筝形与飞镖形的边长又为黄金数比例提供了新的例证。



包含“黄金”比例的飞镖形和风筝形

这些非周期性镶嵌是如此地美丽，而且它们还是新兴的“准晶体”科学中的研究主题。

数中之美

目前为止，我们已经看到了几种特定类型的比例，它们在审美意义上所具有的重要作用，以及由它们所产生出来的一些具体对象。但是在数学内部，人们往往将美理解为内在本质：它可以位于一个困难的问题中、逻辑证明中或是一个壮观的公式中，也可以位于一个令人满意的方程中。而对于上述对象来说，最简单的常常就是最有魅力的。

大名鼎鼎的费马最后定理[①]就是以其简单性而成为一个美丽的问题。这个定理是在17世纪30年代被提出的，它的内容是：对于任

① 17世纪时，费马在对丢番图《算术》的评注中提出了一系列数论命题，但并未加以详细证明，后世数学家陆续对它们进行证明，但最后一条命题的证明要等到20世纪90年代才由英国数学家怀尔斯完成，史称“费马最后定理”。
——译者注

意超过2的自然数n，不可能找到整数值x, y, z使方程$x^n+y^n=z^n$成立。整数6，8，9几乎满足$x^3+y^3=z^3$，因为$6^3+8^3=216+512=728$，而$9^3=729$，但是我们不可能找到使方程左边恰好等于右边的整数。这条定理的美感还在于它历时长久而又难度非凡，它傲然孑立，就好像永远也不会迎来一个证明一样。直到350多年以后，在20世纪90年代，这条定理才被证明是正确的。

下面再来看看数本身。对于思考数学的人来说，数字本身就蕴含着强烈的美感，体现为它们的多样化、它们的排列，以及各种数字序列之中。"斐波那契数列"是由整数所构成的最简单的数列之一，斐波那契（也称为比萨的莱昂纳多）于1202年提出这个数列，把它带入了数学的王国。有一本学术期刊就是专门研究它的，而它的种种新奇性质还在不断地被揭示出来。甚至有人宣称，作曲家是根据这个数列而创作乐曲的。

标准斐波那契数列之美就存在于它的数学之中。要写出这个数列，我们要做的一切就是记住它从数1, 1开始。第三项是它们的和，即$1+1=2$，于是得到数列1, 1, 2。第四项是它前面两项的和，即$1+2=3$，由此得到1, 1, 2, 3。后面每一项都是它前面两项的和，由此我们可以得到斐波那契数列：

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, \cdots$$

你可别以为它只有看起来那么简单。这个数列能够与黄金数φ联系起来。如果我们考虑每一项与它前一项的比值，就会得到下面这个数列：

$$\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{13}{8}, \frac{21}{13}, \frac{34}{21}, \cdots$$

用十进制小数来表示，就变成：

$$1, 2, 1.5, 1.666, 1.600, 1.625, 1.615, 1.619, \cdots$$

这些十进制小数交替地小于和大于黄金比例$\varphi = 1.618\ 033\cdots$的值，如果我们继续取这个比值数列中更多的项，就会越来越接近于φ的精确值。而这种一致性正是数学家们能够欣赏的那种井井有条的美感。

发明中的美

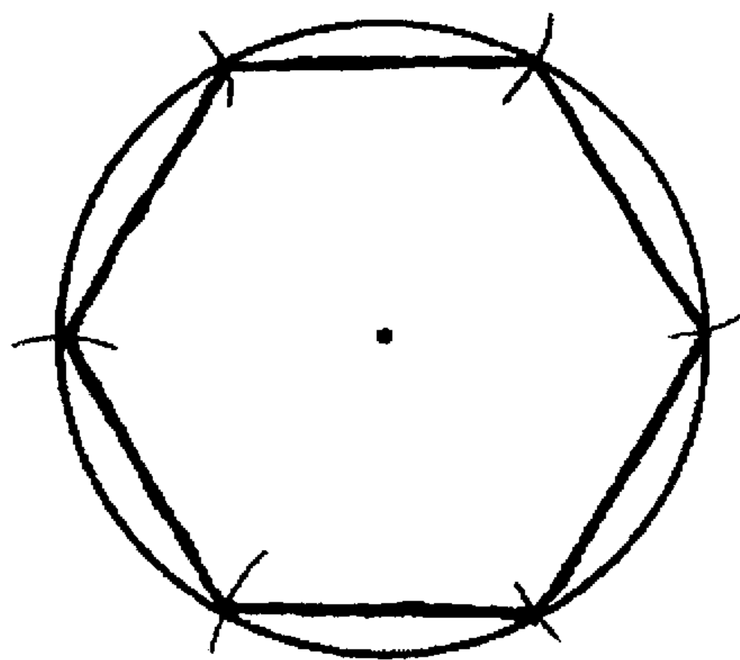
伟大的数学家对于下列问题都会有自己的想法：在他们的发现中，他们自己认为最美的是哪一件？答案并不总是明显的或者最巧妙的那一件。叙拉古的阿基米德生活在公元前3世纪，被公认为古代最优秀的数学家。他对纯数学和应用数学都有贡献，而他对于自己所做工作评价最高的还是球体和圆柱体关系的发现，即，当一个圆柱体恰好围住一个球，两者高度相同且表面相切时，球体的表面积和体积都是圆柱体的$\frac{2}{3}$（参见第11章）。这个不包括公式的发现，以其纯粹的简单性而令阿基米德感到震惊。

> “数学，以正确的眼光观之，所拥有的不只是真理，而且是至高无上的美——俭朴而冷酷，如同雕像拥有的美，没有丝毫人类软弱的本性，没有画作或音乐那么绚烂的装饰，它崇高而纯粹，具有只有最伟大的艺术才能呈现出来的严格的完美。”
>
> **伯特兰·罗素，《数学研究》（1902）**

球体和圆柱体在古希腊几何学中被看做是基本对象。两千多年以后，欧洲大陆的后印象派画家开始将其作为自身之标志，从而变具象派艺术为几何图形描述。一位杰出的后印象派画家——保

罗·塞尚曾建议一位艺术家同行“用圆柱、球和圆锥来看待自然界”，而毕加索和布拉克更是将这一方法带上了新的高度。

19世纪时，受人敬重的卡尔·弗雷德里赫·高斯（他与阿基米德和牛顿一起，同处于最顶级数学家的行列之中）将几种不同的思想结合于他的几何构造之中，由此而得到的优雅而有力的数学证明深为数学家所推崇。他尤其关心尺规作图问题，即，画图时只允许用直尺来画直线（但不能测量长度）和用圆规来画圆或圆弧。



用直尺和圆规画出的正六边形

通过这种方式，不难构造出等边三角形、正方形、有五条等边的正五边形以及有六条等边的正六边形。例如，要构造一个正六边形，我们先选取半径r，将圆规的两脚分开距离r，先画一个圆。用相同设置的圆规在圆周上划分圆弧，经过六次划分后我们回到了起始位置。将分点连接起来，就得到我们想要的六边形。

尽管欧几里得已经熟悉画出这些图形的方法，但是他没有办法构造出具有七条等边的正多边形。高斯从十几岁时就开始研究这个问题，并最终证明这是不可能的。而且构造正十一边形或正十三边形也是不可能的。高斯不是没有找到构造方法，而是证明不可能存在构造方法。他证明了一个否定结果。

高斯的辉煌成就中也包含肯定性的结果。他构造出一个具有17条边的正多边形。实际上他的成就还不止于此，因为17这个数具有特殊的形式。它可以写成如下形式：

$$2^{2^n}+1$$

这是因为，令$n=2$，则$2^2=2\times2=4$，而$2^4+1=2\times2\times2\times2+1=17$。高斯指出，只要这个形式所对应的数是质数，那么具有该边数的正多边形就可以构造出来。具有这种形式的数（称为费马数）在$n=0$, 1, 2, 3, 4时是质数，分别为3, 5, 17, 257, 65 537，但是没有人知道当n取别的值时是否还存在着其他的质数。例如当$n=5$时，费马数为4 294 967 297，可它并不是质数，因为：

$$4\,294\,967\,297=641\times6\,700\,417$$

关于美的问题

由此可见，我们有无数个理由将数学与美联系起来。当然，数学以比例、对称和透视法等形式渗透于音乐、艺术和建筑等领域之中。与上述领域中进行创造的人相比，数学家真的如此不同吗？数学（用自己的语言来讲）也是以美学原理为基础而建立的。即使是对于“旁观之眼”来说，那些拒绝承认数学中的美的人也实在是鲁莽的怀疑者。

16 数学能预言未来吗？
——数学模型、模拟与博弈论

865年，随着美国内战逐渐进入尾声，亚伯拉罕·林肯开始为其连任国家总统做筹划。在就职演说中，他表达了“对未来的高度期望”，但是明智地警示“对于未来之走向没有任何预言”。而正如我们所知道的那样，一个月多一点之后，一个刺客的子弹终结了林肯的生命，尽管他的国家继续走向繁荣。那么，用数学能否帮助他避免这种命运呢？这个问题听起来很荒谬，但是数学的“预测”方法确实已经应用于人类行为与活动的各个方面，所覆盖的领域包括政治、天气预报和股票交易等。

我们很自然地会鄙视那种认为用数学可以预测未来的大胆言论，听起来就好像方程式和数学符号可以变成水晶球一样。自2008年以来席卷全球的金融危机看起来又给数学之自大做出了令人难堪的裁定。若干年以来，非银行金融机构一直在雇用数学家（并称他们为“定量分析家”）来预测未来的发展。但在批评者的眼中，他们那些神秘的数学模型（无论是使用概率还是统计模型）所产生的结果，都是滞后于这场前所未见之金融灾难的。

但是我们要小心，别在倒洗澡水时把孩子也倒出去。从全球银行系统的巨大危机就联想到整个数学建模领域出了问题，这恐怕是错误的。通过很多更为默默无闻的方式，建模方法以极为合理和精确的

结果已做了多年贡献。在再次相信财政顾问之前，我们可能会反复思量，犹豫再三，但是对于在2017年8月21日那场精确到分钟的日食预测，我们却信心十足。

将生活建模

一个数学模型是用数学语言（转译为变量和方程等词汇）对现实生活情景进行描述的一种方式。通过假设什么是重要的以及忽略某些可以不考虑的特殊情况，模型的目标在于把握问题的本质。接下来，我们能够对一种数学机制的建立加以确定，以便考察该模型是否反映了现实生活中的问题。

从传统意义上来讲，数学模型主要应用于物理学中。重要的实例包括，牛顿的引力理论可以对行星运动建立模型，而麦克斯韦方程组（参见第11章）则建立了磁与电相互作用的模型。但是今天，我们可以应用内容丰富的数学技术将很多领域中的问题建模，例如人口统计学、经济学、地理学、心理学、生物学、医学和工程学等。例如，利用这种方式，我们将城市和人口的增长置于数学显微镜之下，从而可以了解疾病在全民甚至在病患群体之中的传播情况。建模的范围还在扩大中。

> “预测是非常困难的，特别是对于未来的预测。”
>
> **尼尔斯·玻尔和其他一些人**

考虑到建模具有极为广泛的应用价值，由此而产生出多种多样的建模技术也就不会让人感到吃惊了，建模总是不可避免地要用到很多种复杂的数学工具，例如概率论、统计学、集合论和“最优化”技术。根据相对论，牛顿的引力理论可能应被看做是“近似的”，但是由它所建立的模型却还是足以预测日月食。在对金融市场进行建模时，不存在严丝合缝的理论——这个例子很好地说明了预测人类的行为比预测行星的行为要更困难。不过在这里，数据仍然存在，

因此借助统计学技术，分析学家还是可以建立方程，设法估计其潜在趋势。

基于计算机与日俱增的计算能力，这个领域现在已经完全建立起来了。但是支撑在技术后面的还是数学理论，它决定着模型在何种程度上反映了现实。

对人口建模

当然，一个模型的效果究竟如何是一个重要而又困难的问题：根据不同问题各自具有的性质，建模问题的最终答案可能在某些时候就是无法证实的，必须等到现实结果出来才行。不过，就在一个领域的已有模型中引入新技术以产生更为精确的预测这个特点而言，人口研究领域倒是一个很好的实例。

19世纪时，牧师托马斯·马尔萨斯在他的《人口论》中提出了一个关于人口增长的著名的数学模型。在书中，他得出了一个悲观的结论：不断增加的人口将会超过土地生产实物的能力。由此他宣称，即使没有流行病、鼠疫和瘟疫夺走大批生命，那么一场“大饥荒”也会导致同样的后果。

用建模的语言来讲，马尔萨斯的理论假设人口每经过25年就会翻一倍，因此，人口N在25年后就会变成$2 \times N$，再经过25年后就会变成$4 \times N$。于是，以25年为时间间隔，人口增长系数为：

$$1, 2, 4, 8, 16, 32, \cdots$$

假定一开始时相对于人口来说有足够吃的食物(F)，马尔萨斯指出，很难想象食物增长的速度能够和人口一样，因为能够用来生产食物的土地是有上限的。所以，他令土地每25年就增加F，于是在第一个25年后可用土地就会有$F+F=2 \times F$，50年后就会有

$F+F+F=3\times F$，以此类推。根据这一理论，食物的增长系数就是：

$$1, 2, 3, 4, 5, 6, \cdots$$

人口的增长系数依几何级数而增加（每次乘以2），而食物的增长系数依算术级数而增加（每次加上1）。马尔萨斯由此得出他的结论：迅速增长的人口将会导致食物短缺。

马尔萨斯的理论在当时产生了巨大的影响，但是以今天的后见之明来看，它明显是有问题的。例如，要用它来预测今天英国的人口的话，就会得到一些奇怪的结果。1800年左右人口约为1000万，按照马尔萨斯所说的，他的理论预测到2000年时将会猛涨到大约2.56亿——是今天实际数字的四倍以上。

公平地讲，马尔萨斯不可能考虑到诸如节育、移民和粮食生产技术等方面的现代发展。但他的模型绝不是白白浪费时间。它促使人们采取更精细的处理方法，建立更合理的假设，从而升级原有模型。现在，我们能够让人口增长率随着人口自身增多而减少，以此而为人口规模设置一个上限（参见第13章）。

预测及其限制

在日常生活的某些领域中，我们所研究现象的变化无常严重地遏制了我们想了解未来究竟会发生什么的欲望。资金流动与天气状况就是很好的例子，尽管我们很努力，但建模方法有自身的限制。

在金融世界中，对股票市场的预测要占据一位职业投资者很多时间。各主要交易中心的股票交易指数（道琼斯、金融时报100指数和日经指数）是通过将一组经过选择的股票合并起来而得到的，而每日交易结束时所发布的则是累加所得的数值。预测第二天的指数可能就意味着收获几百万或是损失几百万的差别。当然，一个众所

周知的困难是，金融市场会受到“信心”这个抽象概念的影响——我们要预测的是人对于那些有时也许看似毫无关联但又可能会对金融活动产生影响的现象的反应。

这时，数学建模始终不变的起点就是历史数据。处理数据的方法才是给出评价的依据。一种简单的股票交易模型，是考察前一天结束时的结果。例如，当我们了解到周一、周二和周三的指数分别是3700、3900和4100时，就有可能以它们为基础来预测周四的指数。在没有其他有用信息的前提下，我们只能求出已有数据的平均数，从而认为我们对周四的预测是：

$$\frac{3700+3900+4100}{3}=\frac{11\,700}{3}=3900$$

这种技术被称为（简单的）*移动*平均值法。之所以用这个名字，是因为在我们得到周四的实际指数后，就可以丢掉周一的数据，而将周二、周三和周四的平均指数作为对周五指数的预测。但是在上涨行情中，平均值法总是给出略有下降的预测，就会因为不能反映出主要情况而被深感不满的交易者所抛弃。为部分补偿这一缺陷，可以将最近的历史赋予相比几天前的历史更高的权重，以便更好地反映大趋势。给每日数据加权方式本身就已经是评价，不同的权重系数体现为不同的模型。当然，在实际生活中，交易者对于一批特定的企业有着专门的了解，并且据此作出全局预测。但是，没有任何预测模型能单凭其自身而预测股市崩盘。

天气预测也会一直是一门不精确的科学（参见第13章）。具有任意程度精确性的预测都只限于对未来的几天而已，但就算是如此，也还是很有可能在实际情况到来的一刻而改弦易辙。困难在于，对大气和海洋环流进行建模并由此而预测天气的纳维尔-斯托克斯方程组，目前仍是无法精确求解的（参见第20章）。我们不得不依赖于

近似解，但是这很复杂，而且涉及大量的计算机计算。

除了天气预测之外，甚至于像多长时间会形成一个长队这种每天都可能见到的简单问题，也会导致那种无法用公式来给出精确解的方程。就算有精确解，它也只适用于满足我们在建模开始时必须提出的那组假设的理想情况。尽管如此，数学家们还是力求通过模拟方法来解决这类问题。

模拟意味着用一种数学模型来进行实验并且进行观测。我们怎样预测要在一天中不同时刻进行操作的超级市场的账目呢？在不同的天气条件下应该如何安排机场上空的飞机分层着陆呢？这几类问题都是用模拟方法来研究的。通过调节模型中的参数（在超级市场的例子中，就是一天中的不同时间；而在飞机的例子中则是“改变”天气状况），做数学实验的人可以使我们看到不同情况下会发生什么，即使是在没有精确解的时候。它并不完美，但是对于凭空臆测和即兴发挥而言，这绝对是进步，而随意的即兴发挥，对于盘旋在机场上空的飞机这个例子来说，则可能导致一场灾难！

关于博弈的教益

一些数学家采用“博弈论”作为进行预测的方法。这种数学理论是由约翰·冯·诺依曼在20世纪40年代开创的，他研究了我们所称的双人“零和”博弈。这类博弈建模的对象是两位玩家之间的竞争局面，两者都会根据已知的支付表来做出理性的选择。该理论引入了特定的心理学假设：参与者是理性的，他们根据最有利于自身这一原则来做出决定。“零和”这个条件意味着如果某一个参与者取胜，则另一方一定会输。这意味着两位玩家之间不存在彼此合作的问题。

也有不假设零和条件的博弈，这就向竞争局面中引入了潜在的

合作问题。这类问题有一个广为人知的例子，就是由阿尔伯特·塔克所设计的“囚徒悖论”问题。其中，两个人，艾伦（A）和布鲁斯（B）因涉嫌拦路抢劫而被警方逮捕，但是警方缺乏足够的证据以指控他们。两位囚徒被关在不同的牢房中，因此没法在一起商讨辩词。两个嫌疑犯坐在各自的牢房中，根据最有利于自身的原则来做出决定，而且他们不能彼此交流以便协调其辩护策略。

艾伦和布鲁斯的“结果”是各自被宣判入狱，但是这不仅取决于他们各自在面对警方提问时的反应，还取决于他们的联合反应——如果其中一人供认不讳而另一个拒不承认，那么警方会“奖赏”认罪者，为他申请比他的犯罪同伙要轻很多的判罚：前者一年监禁，后者十年。如果两个囚徒都招认，那么警方就很方便了，在这种情况下，两人各判四年，全案结束。

艾伦该怎么做呢？如果他招认，那么无论布鲁斯选择怎样的辩词，最大惩罚将是四年监禁。如果他拒不承认，那么最大监禁期限将会是十年。作为一个理性的人，艾伦选择招认，他的策略是将最大可能刑罚最小化，也就是我们所说的“极小-极大策略”。布鲁斯以同样的方式来看待这个问题，因此他也选择了招认——最终结果是两人的下场都是四年监禁。而如果艾伦和布鲁斯待在同一间牢房中，并且能够彼此合作，那么结果就会大不相同。在缺乏足够证据指控的前提下，他们都能够辩称“无罪”，从而以自由人的身份离开。

囚徒悖论问题		布鲁斯	
		招认	不招认
艾伦	招认	4, 4	1, 10
	不招认	10, 1	0, 0

博弈论与核边缘政策

两个小小囚徒的虚构命运只是一个说明性的例子。而在现实世界中，1962年10月间，筹码被无限量地提高了，古巴导弹危机迫使全世界一起面对冷战中核危机的可能性。不过，这个案例也可以用博弈论的方法进行分析，从而成为事件决策基本原理的一个典型案例。1962年，通过军事侦察确认，前苏联在古巴建立了可以打击美国本土的导弹发射场。接着又发现运载导弹的前苏联运输船正驶往古巴，威胁的级别进一步上升了。

用博弈论的语言来讲，时任美国总统的约翰·F. 肯尼迪与前苏联领导人尼基塔·赫鲁晓夫可以有哪些选择？这是一个复杂的局面，两位领导者都不了解对方会做什么。以今天的视角来看，分析这一局面的一种方式是先简化问题，将它看成是一个非零和博弈。美国不得不在封锁古巴和发动空袭之间做出选择，而前苏联也要在撤离古巴和继续停留之间进行选择。可是代价又是什么呢？我们该如何给悲观的结果赋以数值呢？一种方法是用数字将它们分级，灾难性的后果为1级，不够好的后果为5级，从1级至5级，损失程度逐级递减。例如，代价5, 2意味着对美国而言差强人意，但对前苏联而言则非常糟糕，在这个意义上是“美国胜利”。数字表示的是*级别*，在某种意义上也体现出问题的严重程度。

根据实际发生的情况，在那段令人紧张的日子里，两个国家都在努力远离悬崖，尽量避免可能会引发战争的相关行动。美国实施了封锁，而前苏联则同意从古巴将导弹撤离，还有一些让步措施确保最终达成了妥协。

古巴导弹危机		前苏联	
		撤离	停留
美国	封锁	3，3（妥协）	2，5（美国失败）
	空袭	5，2（美国胜利）	1，1（核战争）

现在，博弈论已经可以用来处理超过两个玩家参与的情况，以及允许玩家及小组进行合作与结盟的博弈，例如，允许有两个玩家结成联盟以对抗第三个玩家的情况。对博弈论的研究也有重大进展，美国数学家和经济学家约翰·纳什就这一主题所进行的有巨大影响的研究体现了这一点，他也因此而获得了1994年的诺贝尔奖。

预测：健康警报

不难想到，利用正确、及时的预测是可以避免灾祸的。流行病、海啸、地震和火山爆发都属于这类灾祸，但是我们应对这些威胁的能力是随具体情况而显著变化的。当这样的重大事件发生后，它们的发展是可以描述的，但是要在它们发生之前预测出时间、地点和规模，则如圣杯一般只可想象却难以企及。多数情况下，我们总是满足于一种用概率方式限制其不确定性的预测（参见第10章），例如，“暴风雨降临的概率是70%”，或者“经济增长介于1%~3%的可能性大约为40%”。预测气候变化的棘手问题是体现预测必须以概率为基础的一个重要领域，不同预测之间往往还是有争议的。为气候变化建立模型的人以数学和统计学为基本工具，对自己的预测持有恰如其分的谨慎。只有时间才能告诉我们预测与事实究竟是否吻合。

尽管某些数学模型，例如牛顿的模型，能够给出就其所处背景而言近乎完美的结果，但是其他模型则往往涉及会产生不稳定解的方程。这类模型就要牵扯到混沌理论（参见第13章）。在这样的模型中，利用彼此接近的初始数值可能会得到完全不同的解，而且由于

那些初始数值本身经常就是不确定的，因此预测就变成高度临时性的。

听起来就好像在说：在盲人的国度里，独眼龙就能当国王。在某种意义上讲，这是事实，不过他的视力一直在提高。要想对未来做出成功的预测，关键在于有一种建立在令人满意的科学理论基础上的数学模型。数学和统计学本身作用很有限，但是再结合科学的证据，它们就会变得更有力，目标也更明确。

数学能够预测未来吗？有时是可以的，而且要在正确的背景中，还要有正确的评判。而且有一点好处是我们不该忘记的。无论我们最终能否获得可用的预测，建立数学模型本身仍是一种值得投入的尝试。当我们能够识别变量并且清楚它们相互作用的方式时，就能以更深入的方式思考问题，从而在分析问题时处于更有力的位置。特别是我们对于所使用的假设的有效性会有更好的理解，而要是没有有效的假设，就不可能指望会有任何精确的预测。

17 宇宙是什么形状的？
——拓扑学、流形与庞加莱猜想

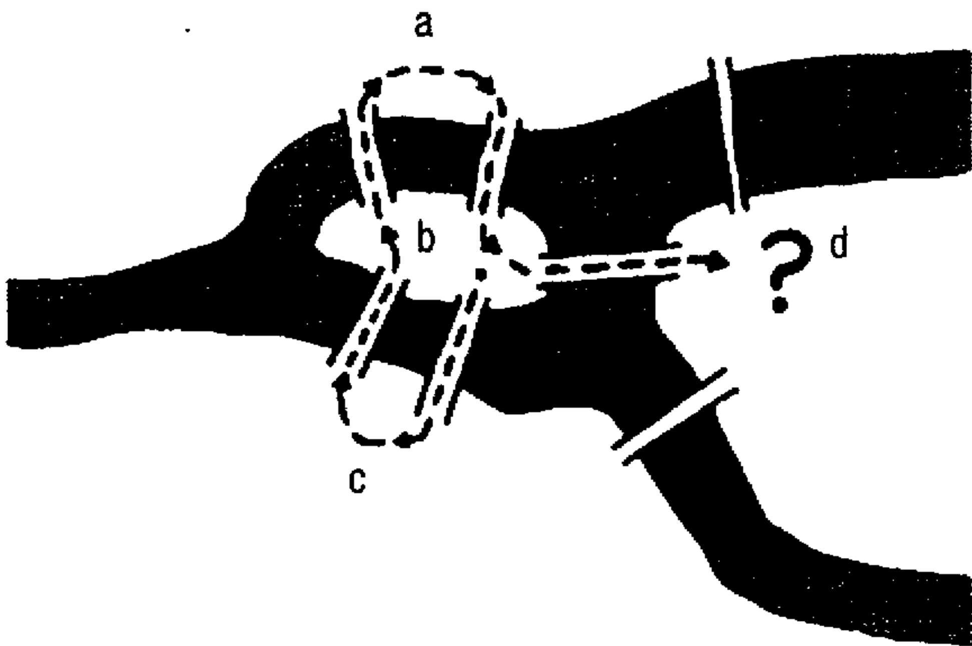
旅行家和远足者经常会与地形学打交道，也就是研究他们所走过的诸如高山低谷等各种突出的特征地貌。数学中也有一个对应的词，叫做“拓扑学”，它也是来自于希腊词topos（“位置”），但它与传统的几何学并不相同。作为一个发展迅猛的新兴数学主题，它需要用到集合论与代数学，而且还取得了令人瞩目的成就。在处理宇宙的形状这一问题时，它遇到了有史以来最大的挑战。

在拓扑学中，数学家们不需要去测量长度、角度、面积或体积。他们更关心诸如各个点如何连接在一起、表面上是否有洞或者图形之间能否互相变换等问题。拓扑学重点关注点集、曲线、图形、曲面和空间的内蕴性质。

建立正确的联系

在伦敦或纽约地铁，或者是巴黎地铁的直观图上，一眼就能看出不同站之间是如何连接的。尽管也是地图，但是其中各站点的分布并不真实反映它们所处的具体位置，从地理学的角度来讲，各点间连线的长度与它们之间的距离没有多大关系。地图只告诉我们怎样沿着铁路网旅行，但没有告诉我们要走多远或者是要花多少时间。它只说明了我们需要知道的东西。

铁路网是拓扑学中的图的一个实例。在拓扑学中，图（或称网络）是一种由“顶点”构成的图像，在这个例子中就是各站，不同顶点之间由“棱”连接起来，这里就是铁轨路线。除了使用不同的术语之外，拓扑学中的图与传统意义上的画在方格纸上可以表示量的变化关系的图像是不同的。在利用顶点和棱的背景下，图讲述了完全不同的故事。



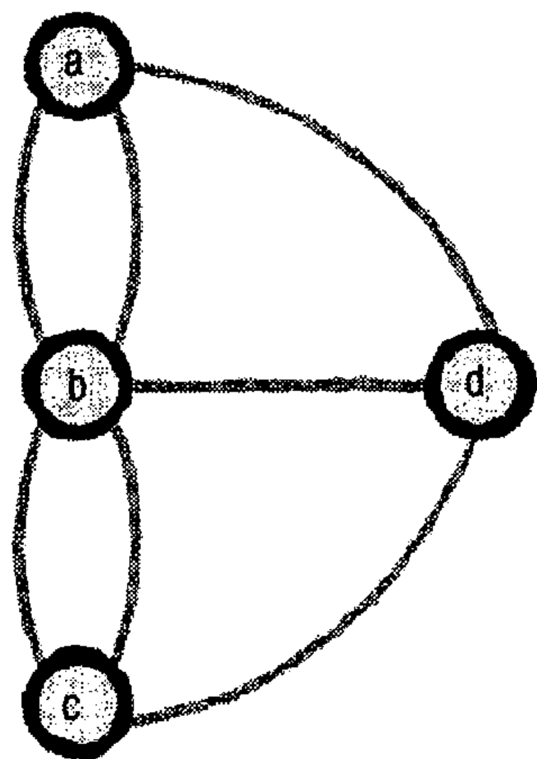
哥尼斯堡七桥问题

人们常常将拓扑学的开端追溯到1735年，当时莱昂纳德·欧拉解决了“哥尼斯堡七桥问题”。是否有可能游历东普鲁士的哥尼斯堡全城（现位于俄罗斯加里宁格勒境内），并且有没有一种方法将所有大桥都穿过且仅穿过一次呢？我们不妨先自己考虑一下。

无论怎样制定旅程计划，我们都会很快遇到困难。我们也许可以经过其中六座桥，但是在这之后，我们不可能在不重复经过已走过的桥的情况下穿过第七座桥。有两种可能的结论。或者是我们的计划制定得还不够合理，或者这是做不到的。这时欧拉前来解围，并且证明了要找到这样的路径是不可能的。但是他真正做到的事情还远不止于这些。他给出了一种判断方法，使我们能够对于任一城市中的任意桥梁分布构型快速判断出是否可能存在这样一条路径。

欧拉证明中的创新之处在于他使用了图论，因为他将实际的哥尼斯堡地图改造成由棱（桥）和顶点（陆地区域）组成的图。他所使用的一个能带来便利的概念是顶点的度数，即“连接”在顶点上的棱的数目，例如b点的度数是5（见插图）。欧拉得到的定理指出：

能在一个城市中的所有桥上只经过一次的条件是，除了最多两个顶点之外，所有顶点都具有偶数的度数。

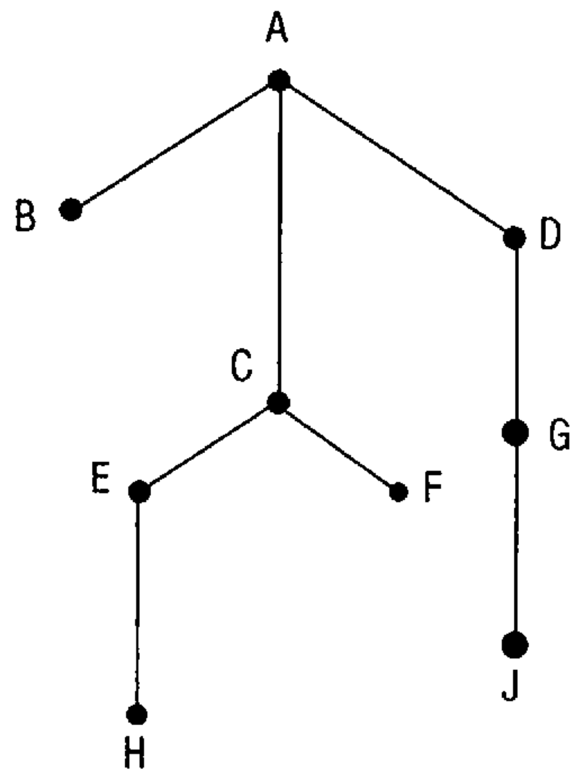


改造成拓扑图的哥尼斯堡七桥

18世纪初叶以后，图论在19世纪大步前行，并在此之后找到了各种各样的应用。今天我们有了万维网，上网者（顶点）与其间数十亿的链接（棱）构成了每一秒钟都在变化的图。

树和它们的根

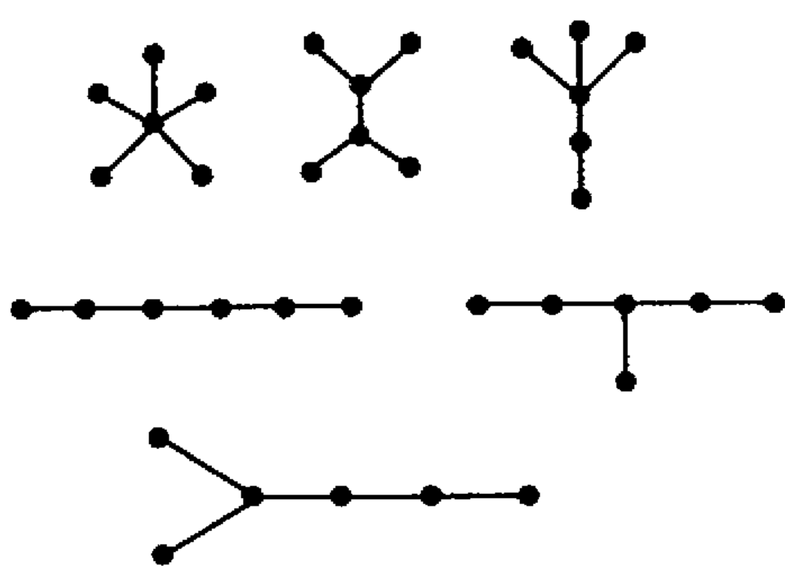
在拓扑学中，树是一种特殊类型的图，在树中从任意一个顶点到另一个顶点都存在着一条路径（这与地铁网不同）。传统的族谱也属于这种类型，而且它是一棵“有根的树”，因为其中有一个根顶点，即位于树顶端的起始点。除了在谱系方面的应用，查尔斯·达尔文也曾选择用这种有根的树来表示他的“生命之树”，以展示物种的演化。有根的树还可以有助于描述族谱，这时我们关心的是（外）祖父母、父母、子女关系，以及第一代表亲、第二代表亲以及更远一代的表亲等。我们还熟悉它的另一种版本：计算机系统中数据存储时所采用的“目录/文件夹/文件”的分级表示。



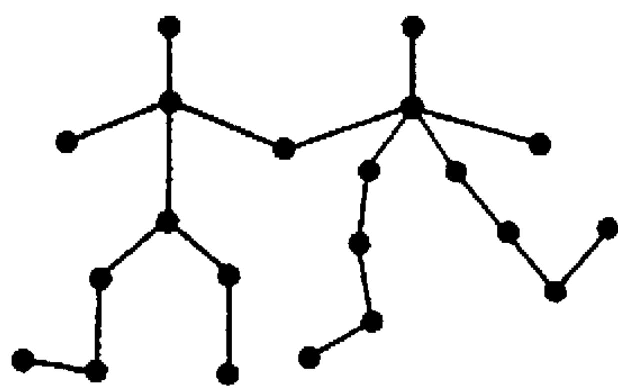
拓扑学中的“有根的树”

不过，也有一些树是没有根的。它们没有优先的根（preferred root），正因为如此也就不存在明确的级别区分。利用给定数目的顶点可以构造出多少无根的树，这个计数问题可谓历史悠久。我们知道，用六个顶点总共可以构造出六个可能的无根树。（通过改变线和点的

方向，还有可能得到看似不同的无根树，可是如果从连接方式的角度来看，只存在六种结构。）但是利用更多的顶点，树的种类将会多得令人震惊。例如，当有21个顶点时，存在着2 144 505种不同的树的结构。



由六个顶点得到的“无根树”



用21个顶点产生的2 144 505种结构中的一种

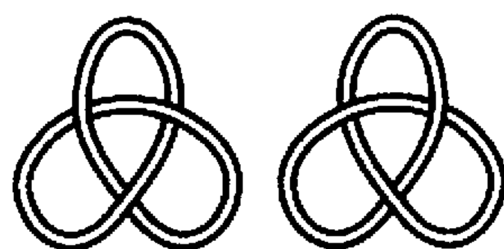
无根树的理论可以应用于有机化学，即用它们来描述分子的结构。顶点代表原子，棱代表原子之间的化学键。重要的不只是分子中的原子数目，还有它们的排列方式及它们的拓扑特征。将同种类型的原子以不同方式排列得到的会是不同的化合物，这种现象称为结构同分异构。

19世纪70年代，数学家亚瑟·凯莱通过计算指出，具有化学分子式$C_5H_{11}OH$的戊醇中的原子一定存在着八种不同的排列方式，尽管当时的化学家只知道其中的两种。自那以后，人们逐渐发现所有理论上可能的戊醇都是存在的。凯莱利用数学的力量对这些化合物的存在进行了理论研究，大大超前于它们在实验室中实际被发现的时间——正如他的朋友J. C. 亚当斯在天文学观测认证之前就用数学方法推导出海王星的存在。这两个实例中的深刻洞见都说明数学有时能够给我们带来强大的力量。

纽结问题

大约在19世纪末时，拓扑学的研究开始超越了顶点和棱，理论

家们开始研究纽结和如何将它们分类的问题。数学意义上的纽结与真实的结并不相同，因为它们没有未打结的末端，因此更像是一根以各种方式缠绕起来的单股线。我们对纽结的数学兴趣在于两个纽结是否是“等价的”，也就是说，是否可以不经过剪断而将其中一个变成另一个。例如，存在两个彼此不等价的三叶草形纽结。



不等价的两个三叶草形纽结

19世纪末时，人们编制出庞大的纽结列表。例如，其中包括165种具有10个交叉的不同纽结。最近，纽结理论被用来研究DNA中的“结”。实际上，DNA是以扭曲缠结形式存在的，但是要先解开缠绕才能使复制更有效，这时数学就有了用武之地。

不动点与多毛球

拓扑学中最重要的成果之一，要从盘子谈起。“布劳威尔不动点定理”这个名字来自于它的创造者——20世纪的荷兰数学家路易岑·E. J. 布劳威尔。让我们从下面的考察开始，将一个盘子放在桌面上，并转过一个角度，那么在盘子的所有点中只有一个没有动，即中心点。这没有什么好奇怪的，但是布劳威尔的不动点定理告诉我们的比这更多，而且具有广泛的应用。

这条定理的内容是：如果我们拿起盘子并将它的边界描画在桌面上，接着便随心所欲地将盘子拉伸、挤扁和折叠，只要保证最后的结果仍处于刚才描画的边界之内，那么就会始终有一个点停留在初始位置上，因此我们称它为“不动点”。这样的点也可能会有很多个，但是布劳威尔告诉我们至少存在一个。关于这条定理，一种富于魔术味道的说法是：如果将一张平坦的纸揉皱后放回桌面上原来的位置，那么纸上始终会有一个点位于揉皱前该点所处位置的正上方。

这个定理可以推广到立体的球，也可以是对于立方体，甚至于多维的物体。因而它很快就广泛流传开来。从事博弈论研究的美国数学家约翰·纳什（参见第16章）利用它证明了有多于两个玩家参与时平衡策略的存在性（由此被称为“纳什平衡”）。在经济学中，不动点定理被用来说明在市场经济中一定的价格对于供求平衡的影响。

布劳威尔定理是一个*存在性*定理的实例。它断言某种对象的存在，但没有更进一步的说明。这类定理属于“定性数学理论”，在这里计算和测量基本上没有起什么作用。另一个拓扑学中的例子是大名鼎鼎的“多毛球定理”。这个定理的内容是：如果我们有一个球，球上的每个点都长出毛发，那么我们不可能通过连续运动将所有毛发都梳理平整，而不剩下至少一根毛发垂直竖立（典型的情况是位于旋涡的中心）。我们不知道究竟哪根毛发会竖起，只知道有一根会竖起。这条定理并非适用于所有几何对象，因为对于炸面圈形，就完全有可能将所有毛发梳倒，不留下任何一根竖起的。

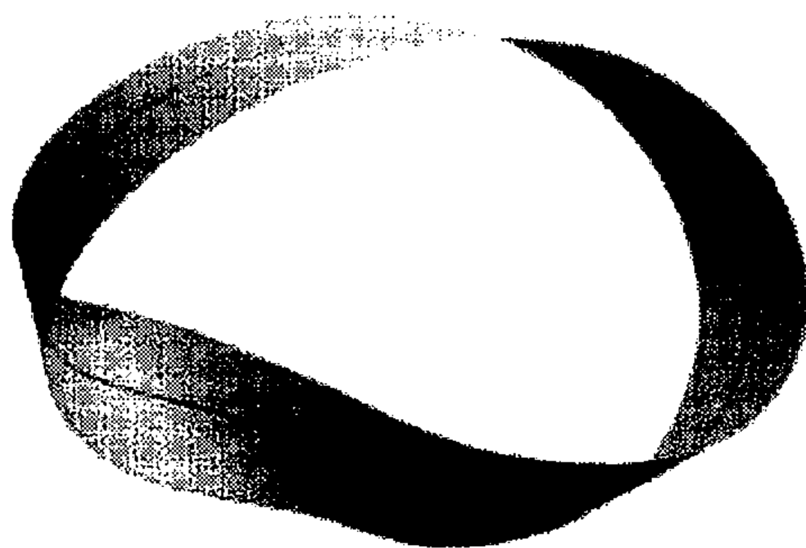
这条看似毫无价值的定理却与对气流模式的研究有关系。我们可以把那个球看做是地球，而把被梳倒的毛发看做是对气流的量度。毛发的方向就是地球表面该点处的气流方向，而它的长度则可以测量风力的大小。多毛球定理告诉我们，地球上一定存在着一个完全没有气流的位置。

流形与莫比乌斯带

“流形”是一种几何对象，它的局部结构，在一维对象的情况下类似于普通的线段，而在二维对象的情况下则类似于普通的圆盘。这种性质可以扩展到 n 维流形。球面是二维流形的一个例子，不过有一个洞的炸面圈形（圆环面）则是另一类流形，我们甚至还可以有多于一个洞的炸面圈形。流形是这样一类几何对象：从整体来看，

它可能会显得很复杂，但是如果只考察它的一部分，看上去就会简单很多。这就是总体景象与局部景象的分别，对于球面来说，我们可以把它理解成是站在一块“平”地上与从太空中俯瞰整个地球的分别。要研究流形的几何，就要用到拓扑学。

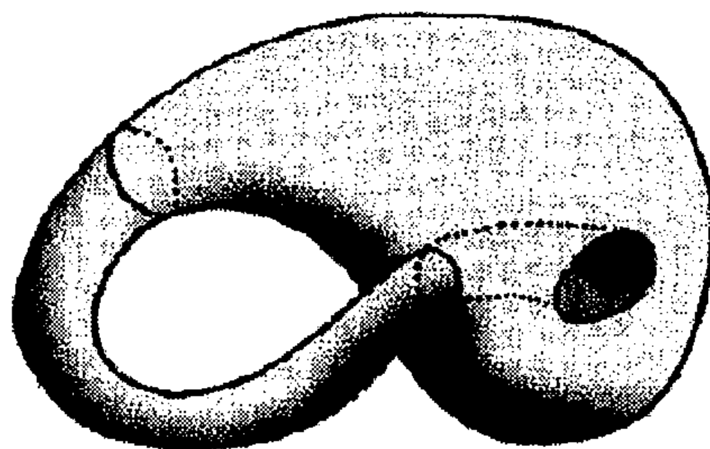
另外，球面和圆环面还是可定向流形。如果我们将（比如说）朝外指向的箭头置于球面上的一点处，并且沿着球面移动这个点，那么它会一直指向朝外，对于圆环面流形，情况也是如此。但是也存在着不满足这种性质的特异的流形。大名鼎鼎的莫比乌斯带就是一个例子：将一张纸带扭转，然后把它的两端粘在一起，所形成的二维表面就是莫比乌斯带。它是单面的，如果将一个箭头置于其上，垂直于表面且指向朝外，然后将它沿纸带滑动，那么它将会回到起点但是指向相反的方向。



莫比乌斯带——一种性质特异的流形

克莱因瓶是不可定向流形的另一个著名例子。它甚至可以用莫比乌斯带构造出来！一首五行打油诗是这样说的：

数学大师克莱因，
钟情莫带视如神。
尝谓莫带云：
两带沿边合，
倏然见吾怪诞瓶。



克莱因瓶

我们需要四个维度才能真正看见克莱因瓶，因此目前能实现的最佳效果

也只能是描画出带交界面的三维图像。

这些各具特色的流形（球面、炸面圈形、莫比乌斯带和克莱因瓶）都是二维表面。例如，一个普通球面是球的边界，尽管球存在于三维世界中，但是球面是球的表面，因此是二维的。

如果我们排除不可定向流形，以及类似无限长圆柱面那样的流形，并且将所考虑对象限于光滑的紧致流形，就会发现任何一个这样的二维流形都可以变形为或者是球面或者是有r个洞的圆环面。由此我们得到了二维流形的分类。

三维流形分类的故事就大不一样了。最简单的三维流形大概就是三维球面。通过与二维球面的类比，它应该是一个四维球的表面，或者相当于四维空间内到原点距离相等的全体点。考虑到我们没办法简单地将四维直观表示出来，在思考三维球面时，我们开始感受到三维流形的复杂性。

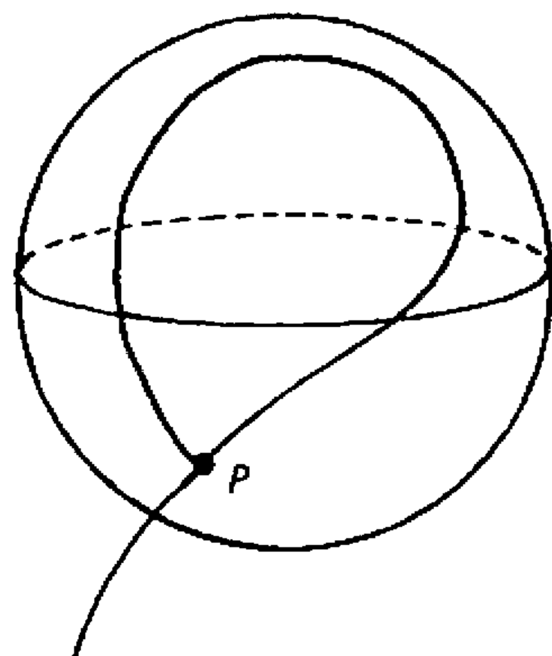
尽管已经取得了一些进展，但是三维流形的分类仍未完成。将很多种不同类型的三维流形划清分界，会得到十分丰厚的回报，因为它们之间的联系，以及在纯数学和应用数学中一些看似没有关联的领域之间的联系，将会因此而变得清晰起来。还有一个巨大的奖励，这个分类结果将会为宇宙的基本形状问题带来启示。因此，三维流形分类的研究纲领具有非凡的重要意义。

在历史上，三维流形与著名的庞加莱猜想有着千丝万缕的关联。

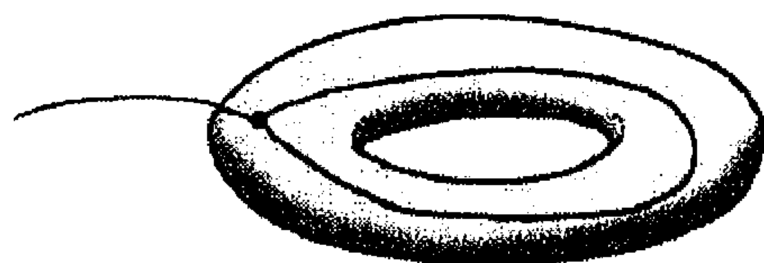
单连通与庞加莱猜想

“单连通的”流形，是由昂利·庞加莱提出的著名猜想中的主题。为了说明这里所涉及的概念，让我们先在一个普通球面上画一个点（P），在该点处附着一个绳圈，然后在球面上移动绳圈。但无

论怎样做，都总是可以将绳圈收拢到这一点，同时保持绳圈与表面接触。对于任意点和任意绳圈，这条性质都是成立的，而正是这一性质使得球面是“单连通的”。相比之下，绕着圆环面上的洞放置的绳圈就不具有这种性质，因为那个洞会阻止绳圈收拢到它所附着的那个点。



一个二维球面是“单连通的”



一个圆环面（炸面圈形）不是单连通的

如同我们刚才已经说明的，由于一个正常的二维流形或者等价于球面，或者等价于有r个洞的圆环面，而在它们之中只有一个是单连通的，那就是球面，因此我们可以得出结论：任意一个二维单连通流形必定等价于一个二维球面。到目前为止这个结论还是很安全的。但是类比于三维情况又会如何？

庞加莱猜想是这样说的：如果一个三维流形是单连通的，那么它一定等价于一个三维球面吗？如果我们能够将三维流形分类，那么就能够回答这个问题，但是这个分类尚未完成。

自从庞加莱在20世纪初提出这个问题以来，它就成为数学中最重要的猜想之一。20世纪60年代，它的一个n维变体，即当n大于或者等于5时的情况，得到了解决，后来$n=4$的情况也得到了证明。仍未得到解决的就是三维即$n=3$的情况。令人振奋的是，隐居世外的俄罗斯数学家格里高利·佩雷尔曼最近将它解决了，回答是肯定的。他所使用的技术与通常拓扑学领域中所使用的技术大不相同，借助了热传导中的数学知识。经过评审，佩雷尔曼赢得了克雷学会

为解决庞加莱猜想而提供的百万美元奖金，但是迄今为止他似乎仍未表露出任何去接受奖金的意向。他就像是无欲无求般地享受着隐遁之乐，视声名如浮云。

宇宙的形状

拓扑学在20世纪初声名鹊起，继而飞速发展，很快就成为一个价值非凡的数学分支。那么它的大问题是什么呢？如果说拓扑学是图形、表面和空间的内蕴性质的几何学，那么宇宙的形状和大小究竟是怎样的呢？

就人类的尺度而言，物理学的宇宙无疑是广袤的。在我们所处的这个渺小的局部中，包含着我们的太阳系，太阳是距离我们最近的恒星，这段距离大约有8光分那么远，也就是，光从太阳到我们这里需要8分钟。第二近的恒星是一颗红矮星——半人马座的比邻星，大约有4.24光年那么远。我们观测到的绝大部分恒星都处于银河系之中，由此构成我们自身所处的星系，而在它之外还有数十亿个类似大小的星系。埃德温·哈勃指出，这些星系正在离我们远去，而且它们离我们越远，移动得就越快。这与爱因斯坦的广义相对论方程经亚历山德罗·弗里德曼修正后而给出的预测是一致的。现在普遍接受的说法是，宇宙创生于大爆炸，而且仍在膨胀之中。但是，它是什么形状的呢？

宇宙形状的问题使我们回想起流形的概念，也就是说，将可观测宇宙的局部景象与整个宇宙的更大尺度的景象相比较。局部几何学会涉及曲率问题，因为根据爱因斯坦的广义相对论，大质量的存在会导致时空弯曲。我们要求助于伯纳德·黎曼所建立的几何学（参见第7章），在这种几何学中曲率会随着位置不同而改变。这就留下了关于整个宇宙形状的大问题，而我们需要拓扑学来帮忙。

物理学家艾德·威滕阐明了拓扑学对于这一问题的重要意义，他注意到“拓扑是当你在未导致断裂的前提下将物体弯曲或拉伸时保持不变的性质”；在爱因斯坦的广义相对论中，时空的结构可以改变，但是它的拓扑是不变的。一般来说，要了解作为四维时空的宇宙的结构，就需要确定描述它的三维流形。

换言之，无论什么时候，只要我们想理解宇宙的形状，拓扑学的理论都是必不可少的。

18 什么是对称？
——模式、对偶性与实在的基本性质

在我们身边，对称无处不在。建筑中有它，花朵中有它，人体躯干中有它。我们很容易识别对称性，因为大脑对于欣赏所见事物中的规则性有一种偏好。自古以来，人类就有一种普遍的共识，认为自然界中的对称性等同于美，典型的例子就是各种令人惊叹的雪花图案。但是如何定义对称性？我们经常会想到“平衡”与“模式”，甚至曾有人将对称描述为“至高无上的均衡”。然而，相比于那些诉诸美学和艺术敏感性的描述，对于数学来说，我们还需要另一种不同类型的语言。

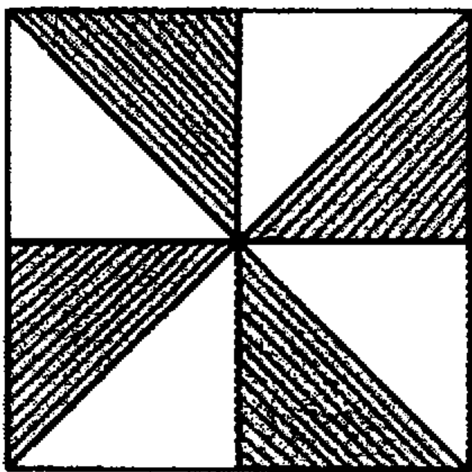
给对称性下一个单一的定义是很困难的。毋庸置疑，存在着那种会使绝大多数人都联想到这个词的性质：成对的观念、对于复制形状的一种感觉及周期性的图案。在数学和科学中，对称拥有各种不同的外观——我们所谈论的是一个多元的概念。

旋转和反射对称性

关于图形的对称性，经常讨论的是旋转和反射对称性。在旋转对称中，我们主要关心旋转轴的位置，经过围绕该轴的旋转使物体，就整个外观而言，保持不变（或者，如同数学家们会说的那样，“使物体变回与自身一致的状态”）。一个二维圆盘就具有这种对称性，一个三维的球也是如此。圆盘和球都具有无限多的旋转对称性。

更有意思的还是那些只具有有限多种对称性的形状，例如，一个带有风车图案的二维正方形。它有旋转对称性，因为如果我们将它按照*逆时针方向*绕着经过其中心的轴旋转（比如说）90°，它将“变回与自身一致的状态”，从而使风车图案的阴影区域再次重合。正方形导致四种旋转对称性，即转过90°、180°、270°和360°（最后一个等价于旋转0°，也就是完全没有旋转）。

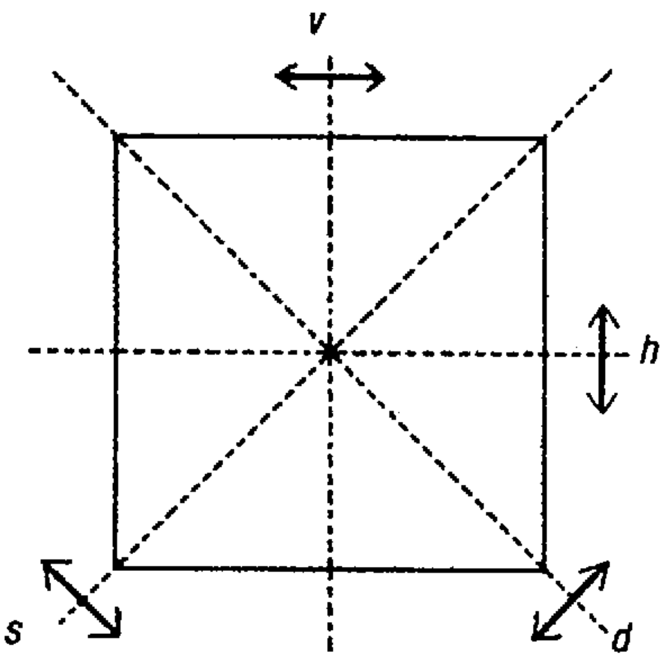
但是，无论是关于它的哪根轴，这个风车正方形都不具有*反射对称性*。不过，要是去掉风车正方形阴影区域，只留下正方形自身，它就有反射对称性了。如果是这样的话，从数学家的角度来说，共有四条反射对称线：一根水平线、一根垂直线、一根“右侧”对角线（从左上角到右下角）和一根“左侧”对角线（从右上角到左下角）。将反射和旋转对称性算在一起，一个空白的正方形总共有八种对称性。



一个带有风车图案的正方形

不变量的对称性

如果测量一下正方形的面积，我们就会发现在旋转和反射之后它的面积是不变的，这与将正方形拉伸的情况是不同的。面积就是*不变量*的一个实例，如果我们只限于考虑某些对称变换的话。



一个空白正方形的八种对称性

由此引入物理学家看待对称性的方式：物理系统在适当对称变换下（例如改变参照系）保持不变的定量或定性属性。因此物理学家是通过不变性来识别对称的。伽利略注意到，一个物

> “对称性（无论你愿意将它的含义界定得宽泛还是狭窄一些）是这样一种观念：在漫长的岁月中，人类一直在努力理解它，并借助它来创造秩序、美和至善。”
>
> **赫曼·外尔**

体，无论是在实验室中还是在匀速行驶的船上往下掉，都会竖直下落，而对称性就在这里。牛顿运动定律的方程就遵循这种对称性。爱因斯坦则更进一步，直接将这种对称性的感觉置于他的狭义相对论的中心，他先验性地宣称光速是一个不变量，而且对于所有参照系都是相同的。

一个物理学家会说，运动方程是变量对应于在不同参照系间转化的变换下的不变量。这样，伽利略、牛顿和爱因斯坦理论中的物理定律就都是对称的。

量度对称性

数学家和物理学家利用被称为*群论*的原理来处理对称性的“度量”和操作等问题。19世纪早期，青年数学家埃瓦里斯特·伽罗瓦关于方程论的研究使群论得以产生。在年仅20岁时便因为一场决斗而过早离世之前，他草拟出整个数学中最美的理论之一——群论的思想（而且长久以来，它的影响力也延伸到物理和化学领域中）。那么，什么是数学意义上的群？

我们还用带有风车图案的正方形为例来说明。回到这个正方形，我们将其旋转90°用符号r来表示，再次应用这个操作，就会得到旋转180°的结果。这种两个旋转的复合，我们用符号$r \times r$来表示，也可以简写为r^2。再旋转90°，则是r^3操作，而r^4则是将正方形旋转360°并回到初始位置。

旋转的复合可以看做是“乘法”，这使我们有可能构造出一张关于旋转对称性的“乘法表”。例如，如果我们想将正方形旋转270°，接

着再旋转180°，那就可以记为$r^3 \times r^2$。这个270°＋180°＝450°的旋转实际上相当于450°－360°＝90°的旋转，在这种情况下，我们可以写为$r^3 \times r^2 = r$。将正方形旋转0°这种特殊的对称性可以记为e，因此我们可以记为$r^4 = e$。

×	e	r	r^2	r^3
e	e	r	r^2	r^3
r	r	r^2	r^3	e
r^2	r^2	r^3	e	r
r^3	r^3	e	r	r^2

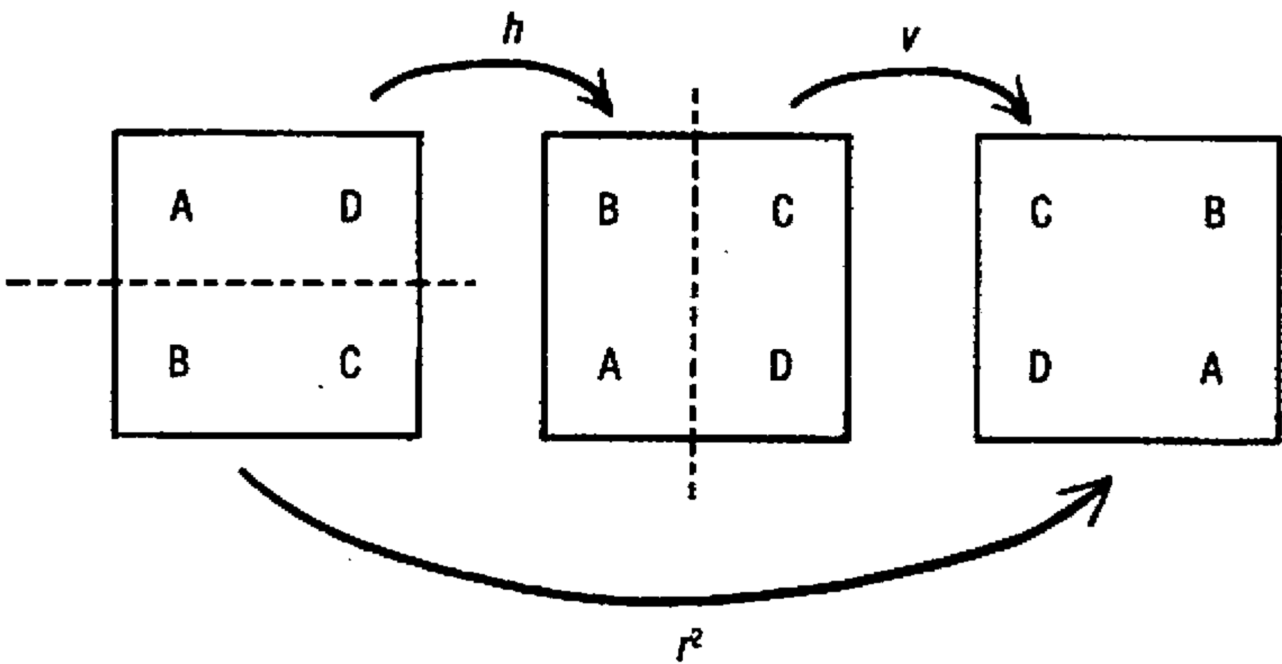
正方形旋转对称性的乘法群表

正方形的反射对称性（水平、垂直、右对角线和左对角线）可以记为h, v, d和s，而且它们能够与旋转对称复合。一次水平反射接着再来一次垂直反射，记为$h \times v$，实际上等价于将正方形旋转180°，因此$h \times v = r^2$。

对于扩充后的“正方形群”，可以构造出一张完整的乘法表，包含反射和旋转八种对称性。

群与对称性

总的来讲，群由“元素”构成，任意两个元素可以复合在一起并且产生一个同种类型的元素。对于对称性的情况，就是两种对称性复合得到另一种对称性。在复合时还需满足某些规则：对于每个元素，都应该存在另一个属于该群的元素与它是互逆的；以及当三个元素复合在一起时，先将哪两个复合对结果没有影响。在这些相当有限的要



反射和旋转的等价关系：$h \times v = r^2$

求之下，大批数学家创造出一种内容极其丰富的理论。

当一个群中只有有限多个元素时，元素的个数就称为这个群的阶，于是我们的空白正方形就有八阶对称性。这个群的乘法表向我们呈现出它的子群，由旋转对称性所得到的子群$\{e, r, r^2, r^3\}$，由水平和垂直对称性所得到的子群$\{e, r^2, h, v\}$，以及由对角线对称性所得到的子群$\{e, r^2, d, s\}$。最后两个群的乘法表具有相同的结果，也就是说，尽管字母表示符号不同，但它们具有彼此相同的乘法表，这种类型的群被称为“克莱因四群”。克莱因四群和旋转群$\{e, r, r^2, r^3\}$是有四个元素的群仅有的两种可能类型。

群论中的一个重要问题是找出具有给定阶数的不同群的数目。例如，4阶群有两个。如果群的阶数为质数，那么这样的群只有一个。如果阶数是合数（不是质数），那问题就有意思多了。例如，有五种不同的8阶群。其中一种是正方形群，如同我们已经看到的那样；另一种是由四元“虚数”（参见第5章），即$1, i, j, k, -1, -i, -j, -k$构成的。

×	e	r	r^2	r^3	h	v	d	s
e	e	r	r^2	r^3	h	v	d	s
r	r	r^2	r^3	e	d	s	h	v
r^2	r^2	r^3	e	r	v	h	s	d
r^3	r^3	e	r	r^2	s	d	v	h
h	h	s	v	d	e	r^2	r^3	r
v	v	d	h	s	r^2	e	r	r^3
d	d	v	s	h	r	r^3	e	r^2
s	s	h	d	v	r^3	r	r^2	e

正方形的二面体群

一种形状的对称性是通过它的对称性群来量度的，因此我们的空白正方形可以通过它的完整的8阶群（二面体群）和它的克莱因四群与旋转群来“量度”。

镜面对称

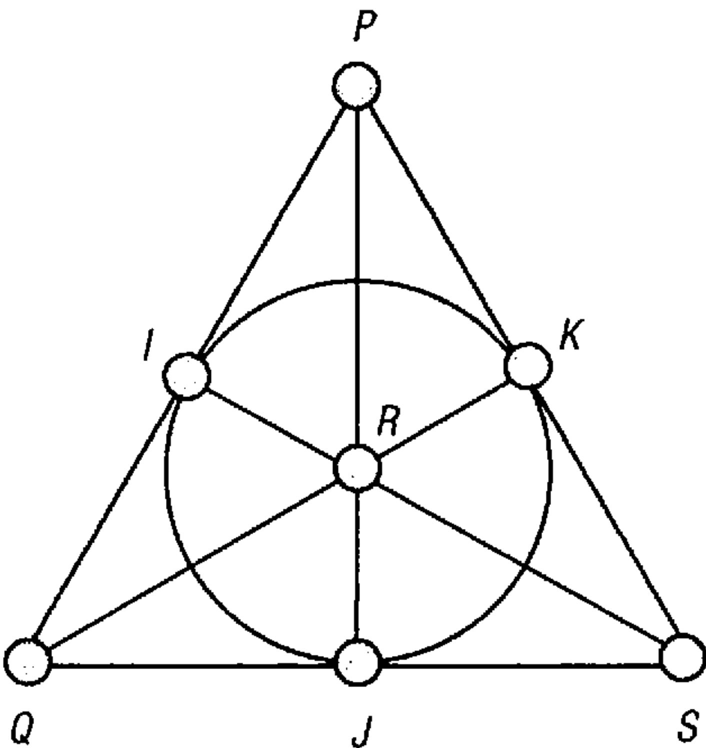
我们可以通过把“平移”和镜面对称性也包括进来而扩充对称性的名单。在考察反射对称性时，我们实际上是在考虑水平、垂直和对角线方向的镜子，但是现在，我们将站在镜子面前。正如所期待的那样，在镜面对称中，我们遇到的是具有自身镜像的物体。当然，镜像中常常会出现反转，这取决于所考虑的对象。将空白正方形放在镜面前，它的镜像中不会有什么明显的不同，因此它具有镜面对称性。但是将一只右手放在镜面前则会产生一个看似左手的镜像。于是，我们将那些在“镜面世界”中具有不同形象的物体称为手性的。在化学中，手性是很重要的性质。例如，柠檬皮中的一种烃类物质的分子，在具有某一种形式时本来是柠檬味的，而它的“手性孪生”分子，叫做柠檬油精，则具有橙子的气味。在物理学中，研究者在探索通常所称的粒子的“弱相互作用”时，为手性的存在而大吃一惊。他们最终得出结论：在建立弱相互作用理论时，应排除镜面对称性。

对偶性的对称

数学家们钟情于对偶和双生的概念，由它们所定义的对称性可以应用于各种符号、定理以及由它们构成的代数知识系统。它的力量体现于用来表达数学问题的语言和名词之中。

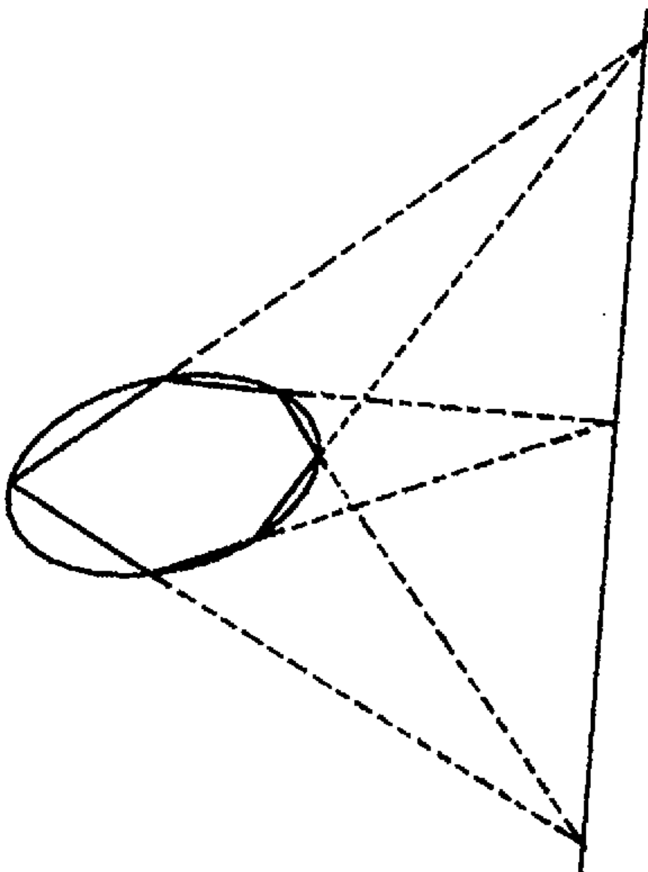
三角形的几何学提供了一个关于对偶性的简明实例。三角形中有三个点（比如说A, B, C）和三条线（比如说a, b, c），点和线之间就存在着一种对称性，因此我们可以说“任意两点确定了一条线”

以及将其中的名词互换后，“任意两条线确定了一个点”。两句陈述中包含着对偶性。这种几何对偶性的用处在于，一个用点和线这两种名词所表示的定理可以自动转化为一个用线和点所表示的定理，两个定理成对出现。



费诺平面

一个内容更丰富的实例是由七个点和七条线构成的几何图形，称为费诺平面（参见第5章）。这七条线中包括一个圆，它被看做是一条线。在这个几何图形中，定理“任意三点唯一确定一条直线”有一条对偶定理“任意三条线唯一确定一个点”。与普通三角形一样，这个图形也是由一定数目的点和线组成的，而对偶性还是会在一般的几何问题中出现。



帕斯卡定理

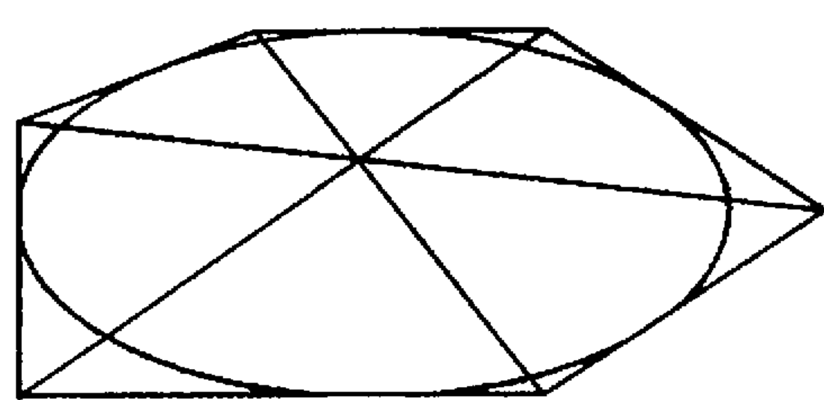
古典几何学中最著名的双生定理当属帕斯卡定理和布里昂雄定理。它们是各自独立被发现的，其间相隔超过150年，因此要通过事后反思才能看出它们是彼此对偶的。帕斯卡定理考虑一个内接于椭圆的六边形，它的内容是：

由六边形对边所在直线相交而形成的三个交点位于一条直线上。

布里昂雄定理[①]的内容是：

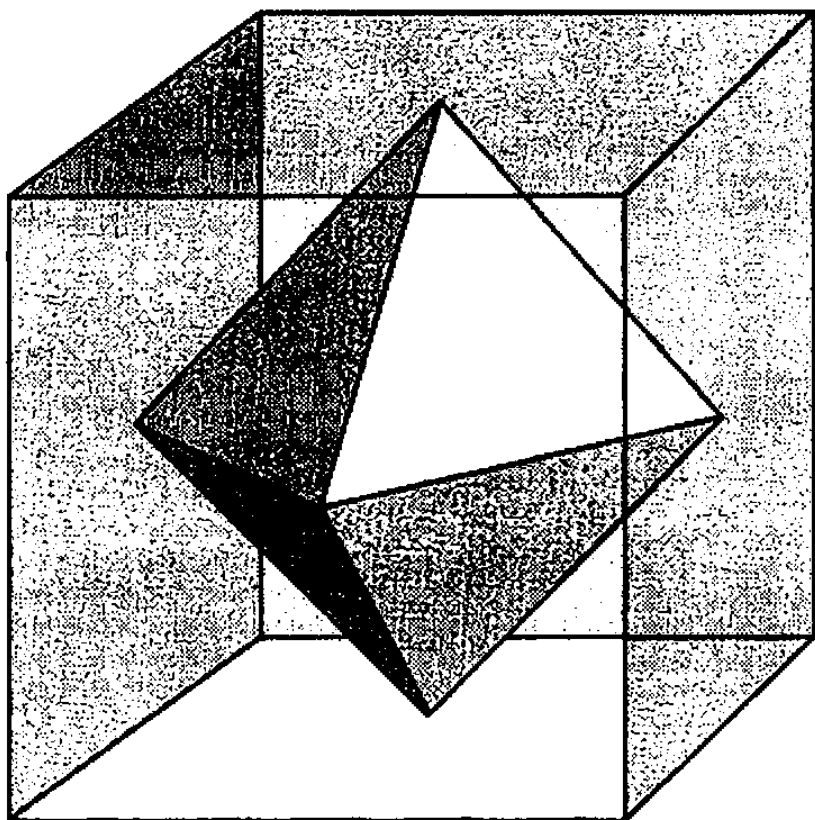
连接六边形对边上切线交点所得到的三条线交于一点。

① 与帕斯卡定理相对应的，这里考虑的是一个六边形的内切椭圆。——译者注



布里昂雄定理

在几何学中，体现为对偶性的对称性是相当普遍的，它们也出现在古希腊人所称的五种柏拉图立体［（正）四面体、立方体、（正）八面体、（正）十二面体和（正）二十面体］中。例如，让我们考虑一个立方体，先找到六个面中每个面的中心点，接着将相邻面的中心点连接起来，所得到的线段构成了一个处于立方体内部的正八面体。换言之，正八面体和立方体是对偶的，这种对偶性质是通过交换面与顶点而得到的。一个立方体有六个面和八个顶点，而一个正八面体则有六个顶点和八个面。类似地，正十二面体和正二十面体也是对偶的，这样就只剩下正四面体，它与自身是对偶的。



立方体和正八面体的对偶性

在逻辑学和代数学中，对偶性也是随处可见，它表现为语言的对称性，以及诸如“OR”和“AND”这样的简单连接词[①]的使用方法。例如，在逻辑学中，“OR”可以用来连接“I have a dog”，“I have a cat”这样的逻辑命题，得到“I have a dog OR I have a cat”。（在日常语言中，“或”通常不包括两者同时发生的情况，但数学中的“OR”也包括这种情况，因此这个复合命题也可以指我同时有两种动物的情况。）

---

① 尽管AND、OR、NOT的含义大致上可以译为与、或、非，但正如文中所指出的，它们与日常语言中的连接词还是有实质差别的。为了强调其作为运算符号的性质，文中将它们直接呈现为符号。——译者注

在逻辑学中，我们将基本命题复合成包括诸如“AND”、“OR”及“NOT”等连接词的更复杂的命题，因此我们可能有：

“I do NOT have a dog OR a cat”

它的意思和下面的命题一样：

“I do NOT have a dog AND I do NOT have a cat”

立 体	对偶性质	面 数	顶点数	棱 数
正四面体	与自身对偶	4	4	6
立方体	与正八面体对偶	6	8	12
正八面体	与立方体对偶	8	6	12
正十二面体	与正二十面体对偶	12	20	30
正二十面体	与正十二面体对偶	20	12	30

五种柏拉图立体的对称性质

在逻辑学中，OR, AND, NOT可以用符号$\vee$，$\wedge$，$\neg$来表示，于是第一个命题就可以记作$\neg(P \vee Q)$的形式，而第二个命题则是$\neg P \wedge \neg Q$（其中P和Q代表单个命题）。它们的等同性引出了德摩根律：

$$\neg(P \vee Q) = \neg P \wedge \neg Q$$

对称性在这里体现为OR和AND之间的对偶性，因此在任何真命题中我们都可以将它们交换。对于德摩根律来说，我们还有另一条，即第二德摩根律：

$$\neg(P \wedge Q) = \neg P \vee \neg Q$$

在集合论中，也可以建立类似的结果（参见第6章）。一个集合是一组对象的总体，而两个集合A, B的并集是，要么从属于这个或那个集合，要么同时从属于两个集合的对象的总体。两集合A, B的交集，则是同时从属于集合A, B的对象的总体。集合A的补集，是不

属于集合A的对象的集合。

用符号∪、∩及^{c}分别表示并集、交集及补集，∪和∩之间存在着对偶性，于是我们在集合论中得到了对应的两条德摩根律：

$$(A \cup B)^c = A^c \cap B^c$$

以及它的“孪生同胞”：

$$(A \cap B)^c = A^c \cup B^c$$

对偶性往往表现为成对出现。而当我们对于对偶结果取对偶时（例如，当我们寻找第二德摩根律的对偶时），就会回到第一德摩根律，这是对偶对称性的又一特征。

亚原子对称性

一线物理学家们正在努力创立一种统一场论，以便建立将自然界四种基本相互作用力联系起来的定律，因此他们已经明确地认识到对称性在这一努力过程中的重要意义。

物理学家们对于“配对”有着自己的理解，这个概念出现在粒子物理学里的一种高度尝试性的理论中，它导致了*超对称*概念的发展，这一概念的背景是通常称为“玻色子”和“费米子”的基本粒子。超对称性的意思是说，对于每一种玻色子存在着一种与之对应的费米子，反之亦然。希格斯玻色子是所有不同类型玻色子中最著名的一种，但我们只是根据粒子物理学标准模型做出推测而已：没有人知道它是否真的存在。人们希望，建于瑞士地下的大型强子对撞机能够解决这个问题。如果这种被称为“上帝粒子”的物质真的出现，那么我们在标准模型中用来解释粒子相互作用的对称性就得到确证，我们也将非常接近于对现实之根本属性的真正理解。

对称的多元性

数学家们不断地与对称的所有形式打交道，并且以此为指导来修正自己的理论。对于他们来说，把握问题的关键往往在于群论，通过群可以定义和探索对称性，进而为欣赏数学美景提供了更为广阔的视野。但是归根到底，究竟什么是对称？也许最好的答案还是一种综合性的回答。它是一系列相似而又各具特色的界定方式的综合体现，从正方形的旋转与倒映到代数学中的对偶性，再到帕斯卡和布里昂雄的孪生定理——不止于此，还深入到（如果物理学家们是正确的话）物质的终极本质。

19 数学是真实的吗？——从柏拉图的实在到哥德尔的不完备性定理

“真理的典范”是诗人柯勒惠支在1791年写给他的兄弟的一封信中用来描述数学的话。他所表达的是古希腊人所持有并且流传了很多个世纪的一种看法，而柯勒惠支本人则是在剑桥大学学习期间汲取到这种思想的。但是那时他正处于一场变革浪潮的顶端，这场变革不仅发生在他自己所处的文学领域内，也发生在数学领域。数学作为“真理的灯塔”的地位即将遭受严格的审查和严峻的挑战。

将数只看做是数的想法无疑是吸引人的：简单直白，不会引起争议（甚至有些乏味）或遭人诟病。我们难道会不同意2＋3＝5是真的并且不可能等于任意其他数吗？三角形内角和是180°而不是181°或者任何其他值，难道不是这样吗？这些数学“事实”自我们很小时便已深深地植根于我们的头脑之中，我们通常不会去质疑其真实性。我们不自觉地认为，其他思想领域可能会因辩论和争执而呈现分裂局面，但是数学似乎在某种程度上却可以超脱出来。这种纯粹性就是柯勒惠支所理解的想法，而且直到19世纪初，它仍然处于正统地位，数以百计的接受大学教育的年轻人在学习古希腊的知识时便接受了这种观念的灌输。

今天，一些数学家解决了高度技术性的难题，或者是构造出处

于数学大厦顶端的深奥理论，他们可能也不会去质疑支撑其成果的数学知识的根基。但是还有另外一些富于哲学倾向的数学家，他们确实在考虑“是什么使数学事实成为真理”这样的问题。他们在数学大厦的地下室中工作，致力于巩固根基，使数学的命题更为可靠。

那么，究竟什么是数学真理的本质？数学家是否也像科学家那样，所发现的只是原先就存在的现象？或者，他们是否实际上更像艺术家，在一张空白的画布上随心所欲地描画？

当我们一开始学数学时，似乎没有什么可以自由处理的余地。每次进行加、减、乘、除运算时都只会得到一个正确结果。这使我们感到，数学在描述现实世界时是一种毫不含糊的语言。因此，数学家的工作就是找出数学真理，利用可靠的原理和无懈可击的逻辑去发现那些“就在那里”的东西，然后将它们收录到数学知识的巨著之中。随着数学真理的积累，我们的知识逐渐增长，记载数学知识的著作也与日俱增。这种确定无疑的论调也同样适用于几何事实，这个领域中有一个持续两千多年的权威，即欧几里得的《几何原本》，这本书从不言自明的真理及定理开始，再利用它们进行各种演绎推理。

真理的探求者

我们有理由猜测，绝大多数数学家将自己看做是原本就已存在的真理的发现者。数学真理的探求者追随着柏拉图的足迹，后者告诉我们，理想形式存在于一个孤立的世界中，一个“柏拉图世界”。他们心目中的数可以看做是柏拉图的形式——稳定不变、内在固有且不受时间影响。

柏拉图在著名的洞穴寓言中，概括了我们在人世间所发现的事物与具有理想形式之真实存在之间的差别。寓言中，一群背对着火把的囚犯坐在洞穴中，他们只能看到存在于洞穴外的真实事物的影

子。洞穴中的阴影构成了囚犯们所能看到的唯一一种类型的实在物，直到有一天，通过获取知识，他们来到了洞穴口，在这里他们察觉到“真正的实在”。

在我们所处的这个不完善的世界中，我们可以画出自己想要的所有三角形，但每个三角形都只是柏拉图式的三角形理想形式的近似。理想三角形应该由没有大小的点和没有宽度的线组成。这个定义是欧几里得告诉我们的，而这是不难理解的，因为欧几里得曾经在雅典向柏拉图的传人学习数学。在《几何原本》中，欧几里得阐述了关于有两条相等边长的三角形（即等腰三角形）的定理，证明它们具有相等的底角；但是在这样做的时候，他所证明的事实并不是针对我们可以在书中看到的图画，而是柏拉图世界中的“实在”三角形。按照柏拉图的说法，只有“用心灵之眼”才能看到实在之物。

柏拉图哲学对于其后各个时代的数学思想产生了强有力的影响。它促进了数学探索过程中对于精神层面的追求。英国数学家哈代十分认同这个观点，他在《一位数学家的辩白》中这样写道：“我相信数学的实在性存在于我们之外，而我们的作用只能去发现或观察它，那些为我们所证明的定理，或者被我们大言不惭地描述为自身之‘创造’的东西，仅仅是我们对于所观察到的结果的记录。”

学生：于是，一切算术和计算都必须用数来进行？
教师：是的。
学生：而且，它们似乎正引领着我们的思想通向真理？
教师：是的，以一种极其不寻常的方式。
学生：那么，这就是我们所要寻找的那种类型的知识。

一段柏拉图式的对话

真理的创造者

到19世纪末，当数学家们重新展望未来目标时，数学家成为数学知识的创造者与发明者而不仅仅是发现者的可能性增加了。数真的存在吗？从一方面来讲，显然不是，因为它们只是语言的符号和表现形式。由此更进一步，可以说我们创造了自己需要的数学。这种想法导致了与柏拉图传统截然不同的数学哲学。

德国人利奥泼得·克罗内克有一句名言："上帝创造了自然数(即整数)，其余均属人为。"克罗内克想要把一切都还原成算术，从而最终回归到整数的层面。他坚信一切数学推理都必须由有限多个严格步骤组成，而且他只承认能够从整数出发以这种方式构造出来的数学对象是存在的。基于这个理由，对于克罗内克来说，常数π是不存在的，因为我们不可能了解它的无穷级数展开式中的所有项：

$$\pi = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$$

在费迪南德·冯·林德曼证明π是一个超越数（参见第4章）之后，克罗内克称赞他得到了一个漂亮的证明，但同时又说，这并没有证明任何事情，因为π本来就不存在。这是对于自古希腊时代以来所持有的数学传统观念的一种冲击，因为π一直被认定是一个与圆有关的数，而这就足以证实其存在性。

在19世纪，我们目睹了数学学术研究的飞速发展，这体现为大学中数学教授职位的创设，数学学科的普遍职业化，以及由此导致的新期刊的大量出现。这些效果累积起来，将数学发展中的涓涓细流汇成滚滚洪涛。新发展不断对关于数学只是对"真理"之发现的观点提出质疑，从而产生对数学根基的重新评价。数学家们开始研

究能够使$a \times b$与$b \times a$得到不同结果的奇怪的新型乘法（参见第5章），从而使代数学走上了新的发展方向，还提出了与欧几里得不同的新型几何学（参见第7章）。

> “一切数学都只是符号逻辑，这一事实是我们这个时代最伟大的发现之一；而一旦确认了这一事实，那么关于数学原理，剩下的就只是对于符号逻辑自身的分析。”
>
> 伯特兰·罗素，《数学原理》（1903）

数学家们对于数学这门科学及其界定开始有了越来越明确的自我认识。美国数学家本杰明·皮尔斯提出这样的观点：“数学是导出必要结论的科学。”在皮尔斯心中数学和逻辑是密不可分的，而且他并不是唯一一个这样想的人，伯特兰·罗素在《数学原理》（1903）中对于纯粹数学有一段著名的描述，称数学是“一切形如‘p蕴含q’的命题所构成的类”。对于罗素来说，重要的是这个逻辑结构而不是p和q的实际真实性，这种想法促使他后来将数学概括为“这样一门学科，在其中，我们既不知道自己在谈论什么，也不知道自己所谈论的是否真实”。

新的方向

所有这一切对于数学来说意味着什么？在某些情况下，这是新柏拉图主义者与激进分子之间的一场较量，曾经确凿无疑的“真理性”变成了令人困扰的一系列人造物，甚至出现了几种竞争性的真理。在德国，康托尔创造了一种集合理论，它有可能会为数学提供一个合适的基础。

集合是若干对象的总体。在数学中，比如我们想了解关于整数的某些事情，就会谈到整数的集合。它看起来像是一个不会带来什

么麻烦的概念，但是罗素几乎只是一击便差点将它扼杀在摇篮中，他从中导出一个内在矛盾，也就是我们今天所熟悉的罗素悖论。

为了考察罗素悖论，我们需要这样一个集合：它可以作为自身所包含的一员。让我们先来考虑一个集合，比如说集合A，它是由一切抽象事物组成的。集合A作为全部抽象事物的总体，本身也是抽象事物，因此A也是自身中的一个成员。这种关系可以记为$A \in A$，其中符号“$\in$”表示“从属于”。但是如果我们现在来看看集合S，它被定义为由所有不是自身成员的集合所组成的集合，就会得到罗素悖论。其中的道理是这样的。让我们先假设S是S中的一个成员（即$S \in S$）：那么从逻辑上讲它就必须满足集合S所定义的关系，但是那就意味着S不是自身的成员。数学家们将这一结果记为$S \notin S$。但是，如果我们从$S \notin S$开始，那么根据这一性质，S确实满足S中成员的定义，因此我们又回到$S \in S$。

结果就是罗素悖论：S是它自身中的成员，当且仅当它不是自身中的成员！它与广为人知的理发师悖论相似，后者的内容是：村子里的理发师得到通知，要他给村子里所有不给自己理发的人理发，那么他应该为自己理发吗？如果他给自己理发，那么他就不该给自己理发；而如果他没有，那么他就应该这样做。在任何一种理论中，这种固有矛盾都是要避免的。

克服这一困难的一种方法是，构造一组集合论公理，使那些会导致悖论的集合成为不合法的集合。其中一组公理是由恩斯特·策梅罗和亚伯拉罕·弗伦克尔提出的，后来又于20世纪的前10年中将“选择公理”补充进来。一般来说，大家都认可这是一组很有用的公理，它成功地阻挡住悖论的侵扰。但问题依旧存在。由这组公理所界定的集合论系统是否就不再有矛盾了呢？是否还有别的没有预料到的悖论在伺机而动呢？

"……数学中存在着相当多的争议。纯粹的数学家拒绝承认应用数学家的证明，而逻辑主义者也不肯承认纯粹数学家的证明。逻辑主义者对形式主义者的证明不屑一顾，而某些直觉主义者则满怀轻蔑地排斥逻辑主义者和形式主义者的证明。"

埃姆雷·拉卡托斯，数学《科学与认识论》

形式主义者

为了给数学提供一个坚实的基础，德国数学家大卫·希尔伯特提出，应该以公理化形式重新处理"古典"数学，因而可以从公理出发，经过有限多步论证以合乎逻辑的方式演绎出各种定理。2000年前，欧几里得为了证明几何定理曾经使用过同样的方法，但在那时，欧几里得将公理或"公设"看做是自明的真理。而根据希尔伯特的形式主义原则，符号与意义是割裂的，我们只能根据公理来操作符号。符号不一定要代表任何实际意义，为了说明这一点，希尔伯特指出，即使将"点和线"等名词换成"桌子和椅子"，也还是可以通过演绎来得出几何学中的各种结果。

如果我们采纳这种"形式主义的"方法，那就不难想象一种编程计算机，它利用公理和逻辑规则，通过有限多步推理，演绎出各种定理链。一切的关键在于，符号操作必须在公理和规则限制范围之内。在欧氏几何学中，定义点和线是件颇费思量的事情——点是没有大小的东西，线则是无宽度的延展。但是在形式主义（它对于现代数学的塑造有巨大的影响）的处理中，我们不需要那样严格地定义数学对象。希尔伯特对形式主义有很高的期望，盼着它能为数学提供一种令人满意的规范。

直觉主义者和不完备性定理

荷兰人路易岑·E. J. 布劳威尔对于希尔伯特与罗素的看法都持反对意见。他排斥按照罗素的观点将数学等同于逻辑，也不相信仅仅通过设定数学对象应该遵循的一组公理就可以真正理解数学对象。布劳威尔认为，数学优先于逻辑，而数学对象则是通过“直觉”来认识的。由于认为数学对象是由直觉指导的，布劳威尔和他的追随者们强烈质疑对无限集合的处理方法。特别是，他们拒绝接受“排中律”——一条自古希腊时代流传至今的逻辑原理。排中律指出，任何命题要么自身为真，要么它的否定为真。希尔伯特等形式主义者对于直觉主义者的立场感到惊惧，因为他们意识到，如果丢掉排中律，那么一直以来安放在数学知识宝库中的很多古典数学重要成果就都要遭到抛弃。类似欧几里得对于质数有无限多个的证明那样的美妙定理也将离我们而去[①]。

> “哥德尔的定理意味着纯粹数学是永无穷尽的。无论我们解决了多少问题，总会存在着无法在现有规则内解决的其他问题。”
>
> **费里曼·戴森**

希尔伯特希望将数学奠基于形式化公理系统的理想在1931年遭受了致命的打击，这一年奥地利数学家和哲学家库尔特·哥德尔证明了“不完备性定理”。哥德尔证明，在一个描述自然数1, 2, 3, 4, …的算术知识的形式系统内，存在着既不能被证明也不能被推翻的命题。因此这些命题仍旧是不可判定的，而该形式系统则是不完备的。在随后的一个定理中，哥德尔证明，如果一个公理系统是无矛盾的，那么它的无矛盾性是不可能在系统内部得到证明的。

① 从第3章中提供的论证过程看，这个危险似乎很有限，因为从任意有限多个质数出发，总可以构造出更多质数，于是质数的数目必定可以超过任意给定的量。——译者注

哥德尔向我们表明，数学真理性这一观念一定会遭遇困难，因为在形式算术系统中存在着这样一些结果，它们不可能通过有限多个类似计算机那样的步骤来证明。这与希尔伯特关于真实陈述必定可以证明的信念是违背的。哥德尔的成果所导致的一方面影响是，我们可能会开始相信数学中某些困难问题其实就是不可解的。不过我们还知道，尽管可能需要几百年的时间，但是确实有某些问题最终得到解决：我们都亲眼目睹了费马最后定理（参见第15章）和庞加莱猜想（参见第17章）的证明。也许我们能从其中获得的真正教益在于，人类的头脑所具有的灵巧力量最终总是可以超越机器的计算能力。

数学的真理性——自相矛盾？

那么，我们的数学真理之书究竟情况如何？哥德尔告诉我们，它并不像最初所认为的那样安全。出生于匈牙利的哲学家埃姆雷·拉卡托斯用另一种方式指出，根本不存在像确定的数学知识那样的东西。相反，随着用来描述数学定理的概念不断改变含义，定理本身也在不断地变化着。

到了20世纪，当我们追问“数学是真实的吗？”这个问题时，我们所考虑的内容也已发生了变化。借用理查德·柯朗和赫伯特·罗宾斯在他们那里程碑式的著作《什么是数学？》中的话来讲：“不再追求诸如理解‘事物之本质’或是掌握‘终极真理’这样的目标，是一个了不起的进步。”哥德尔指出逻辑的完美性只是幻梦，但是在某些方面，这恰好使数学家拥有探索的自由。真理，一如美，可能就在旁观者的眼中，这个道理有着解放性的一面，于是我们可以说，今天的数学家们在他们所能看清的范围内追求着真理。

20 还有什么未解之谜吗？
——未解决的大问题与数学的未来

现于在校学生练习本上的那种板上钉钉的答案，无论最终得到的是对勾还是叉子，可能会使人觉得数学在总体上就是界限分明和固定不变的。当然，事实远非如此。仍然有一些声名远扬却令人无计可施的难题遗留至今，困扰和挫败了一代又一代试图去解决它们的人。而每个已解决的问题又可能会引发数百个新问题，使数学始终保持着旺盛而蓬勃的生命力。

数学的成长要靠未解决的问题，这样的问题还真不少。其中不乏有数百年历史的老问题，而新出现的理论也会带来新的挑战。数学家们会周期性地列出重要而未解决问题的“采购单”，最著名的实例当属大卫·希尔伯特的“23问题”，那是在1900年时作为给新世纪数学家的挑战而提出的。其中一些是具体的问题，还有一些则是开放性的规划。由于希尔伯特在数学世界中所具有的崇高地位，它们吸引了大量的关注，这23个问题在过去一个世纪多一点的时间里照亮了数学发展的前景。它们成为世界上最出色的一群数学家集体行动的焦点。

作为激励，一些高额奖赏出现了，悬赏给那些有能力对某些困难问题给出证明或解答的数学家。最早出现的高额奖励之一要算沃尔夫斯凯尔留下的100 000马克奖金，保罗·沃尔夫斯凯尔这位德国大亨奖赏的目标是费马最后定理（参见第15章）的第一个有效证明。

这个特殊的定理最终于1994年由安德鲁·怀尔斯爵士所证实。在英语世界中最广为人知的奖赏可能要数克雷数学研究所（一个致力于推广数学知识的私人基金会）提供的奖金。克雷数学研究所为经裁定对于21世纪数学发展极为重要的七大问题的解决分别提供了100万美元。在这七大挑战中，关于几何对象之拓扑研究的庞加莱猜想已经得到了证明（参见第17章）。

在遗留至今的这些难题中，每个数学家都会有他所钟爱的谜题。不过一般认为下面这几个问题堪称其中精华：哥德巴赫猜想、黎曼假设、纳维尔-斯托克斯方程（组）以及简称为$P=NP$的问题。值得对它们逐一加以考察。

哥德巴赫猜想

哥德巴赫猜想是18世纪中期由克里斯蒂安·哥德巴赫提出并由此而得名的。希尔伯特的列表中也曾谈到过它，但是所有试图证明它的努力都还没有奏效。数论是一切数学知识的核心，它产生出一些极为困难的问题，其中最有代表性的就是哥德巴赫猜想。数论问题具有容易叙述但难于证明的特点。

作为这一类型的众多问题中的一个，哥德巴赫猜想考虑的是将数表示为质数之和：

每个大于或等于4的偶数都可以写成两个质数的和。

为了检验，让我们先随机选取一个偶数（比如说407 308）来确认一下。事实上我们会发现，407 308是两个质数的和：17＋407 391。目前还没有人能给哥德巴赫猜想找出一个反例，但是也没有人能够证明它，因此它仍旧是一个猜想。数值结果表明，直到10^{18}为止它都是正确的，也就是说在小于该数的范围内是不可能找到反例的。

黎曼假设

哥德巴赫猜想诚然是一项艰巨的挑战，但是在数学意义上，大名鼎鼎的黎曼假设才是重头戏。作为希尔伯特列表中的第八个问题，它是一切数学知识中最迫切需要证明的，任何成功证明它的人必将获得巨大声望——还有100万美金。数论中的很多结果都取决于它的真实性，因此它从不缺乏数学侦探们的关注。

黎曼假设可以借助于通常所称的“黎曼zeta函数”来进行描述。这个假设源于19世纪中期的伯纳德·黎曼，它的名字来自于希腊字母表中的第六个字母 ζ（读作zeta）。这个函数可以表示为：

$$\zeta(s)=1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\cdots$$

对s选取一个数值，可以计算函数的值，例如当$s=2$时，这个函数的值是：

$$\zeta(2)=1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots$$

乍看起来，很难想象为什么这个级数会与常数π扯上关系，但是由莱昂纳德·欧拉得到的一个著名结果告诉我们，$\zeta(2)$的值是$\frac{\pi^2}{6}$。

关于黎曼zeta函数，一件颇为不寻常的事情是它与质数之间的联系。欧拉发现，这个函数等于所有形如$\frac{p^s}{p^s-1}$的数的乘积，其中p是质数。例如：

$$\zeta(2)=\frac{2^2}{2^2-1}\times\frac{3^2}{3^2-1}\times\frac{5^2}{5^2-1}\times\cdots=\frac{4}{3}\times\frac{9}{8}\times\frac{25}{24}\times\cdots$$

在那篇著名的1859年论文“论小于给定值的质数的个数”中，

黎曼引入了令s为“复数”即二维数（参见第4章）的可能性。

复数具有$a+b$i的形式，其中a, b为实数（i是“虚数”，参见第5章），接着黎曼开始寻找那些能使$\zeta(s)=0$的s。前三个值近似为：

$$\frac{1}{2}+14\mathrm{i}\text{、}\frac{1}{2}+21\mathrm{i}\text{和}\frac{1}{2}+25\mathrm{i}$$

这样的值有无限多个，但是已经计算出来的那些（已有数十亿个）都是以$\frac{1}{2}$为实部。黎曼假设的内容是：所有使$\zeta(s)=0$的值都具有这种性质。目前我们已经证明有无限多个值具有这种性质。但问题在于，从逻辑上讲，仍然可能有无限多个解不具有这种性质。

黎曼指出，这个假设与质数在数直线上的分布方式有联系。质数在（比如说）前100个自然数中分布比较稠密，但是沿着数直线展开后，在某些质数间会出现巨大的空隙。黎曼假设掌握着了解这种分布的具体规律的关键。

纳维尔-斯托克斯方程（组）

当我们乘飞机飞行或坐轮船航行时，可能会不幸地遭遇到空气和水的湍流。水和空气都被归为流体，数学家和科学家研究它们的性质已有几个世纪。在19世纪前半叶，法国物理学家克劳德-路易斯·纳维尔和出生在爱尔兰的数学家乔治·盖布瑞尔·斯托克斯对流体进行了彼此独立的研究。今天，湍流的数学模型就是靠纳维尔-斯托克斯方程（组）来建立的，方程也是依据这两位学者的名字来命名的。

这个方程组是根据物理定律（质量守恒定律和动量守恒定律）推导出来的。麻烦在于，到目前为止我们还是不会精确地解它们，而且在用来解释方程解的性质的数学理论方面，几乎毫无进展。如

果能够取得进展，那么天气预报就会不那么像艺术而更像科学，其他一些领域也会因此而受益。烟的卷缩运动和火焰的形状也呈现出与空气湍流运动类似的物理特征，因此这些现象也可以用这个方程组来建立模型。

纳维尔-斯托克斯方程的求解堪称数学世界中的圣杯，它也是克雷数学研究所提出的七大挑战中剩余六个未解决难题中的一个。要赢得数学研究所提供的奖金，需要对方程组的性质产生实质性的洞见，从而建立一种全新并且深刻的流体动力学的数学理论。

利用近似的数值方法已经取得了很多进展，现在，一个称为计算流体动力学（CFD）的全新领域已成为当前研究中一个非常活跃的方面。与CFD相结合，计算机制图使我们能够将湍流可视化进而建立模型，特别是应用于飞机和轮船设计。计算机制图可以用虚拟实验来替代物理的风洞实验，这些虚拟实验具有更显著的实用优势，例如，它们可以实时模拟机翼上方的气流与经过船体的水流。这意味着我们可以调节虚拟飞行器或船只的速度，并且看到由此产生的直接影响。这可不只是计算机动画而已，因为流动模式是直接来自于方程的。随着计算机速度的增加，这样的模拟将会变得越发富于现实感。

P＝*NP* 吗？

*P*是否等于*NP*呢？这个看似神秘的问题处于计算机科学与数学交汇的所在。作为“计算复杂性理论”的一部分，它所讨论的是计算机处理能力的极限。它也在克雷数学研究所那张关于解答与证明的“最高通缉”列表之中。这个问题在21世纪中显得尤为重要，因为它对于计算机安全领域和数论中所使用的算法问题都会产生影响。计算机的工作必须依赖于算法，即一系列需要执行的指令，某

些算法只需要几微妙就可以实现，而另外一些，以目前的速度来看，则可能需要几十亿个世纪。

算法的效率是一个至关重要的问题。例如计算机经常要做的一件事是排序，按照字母表顺序排列文件名称，或者是按照降序来排列数字。例如，一个人要将数字5, 3, 4, 2, 1按照正确顺序排列，这项工作还是相当简单的：我们知道应该以什么顺序排列。但是一台计算机怎样处理这样的任务呢？一种常用方法是“冒泡”排序法，该算法的过程是，按照一种有规律的方式考察相邻的数对，如果其中一个大于另一个，要么将它们交换为所需的顺序，要么当它们处于正确顺序时保持不动。

在对该序列的第一轮考察中，计算机首先将5和3交换，得到3, 5, 4, 2, 1。接着看第二对数，又将5和4交换，得到3, 4, 5, 2, 1，依次类推。经过四次这样的比较后，第一轮结束，数5成功地“冒”上来，到达它的正确位置即序列末端。在下一轮中，计算机只需要考虑3, 4, 2, 1即可。累积起来，为了得到准确的序列，计算机一共要进行十次比较。

推而广之，我们可以说，如果有n个数要进行排序，那么我们可以用同样的方式来计算比较的次数。对于所需步骤数存在一个界限，因为很明显它们总共也不到n^2(正如我们在10小于$5^2=25$这个例子中所看到的)。任何步骤数具有n的幂次形式的算法称为可以在“多项式”时间内完成的。计算机可以轻松地处理这类问题，这样的算法就是高效率算法。

现在让我们来考虑一下，计算机可以怎样处理著名的“流动销售员问题”。在这个例子中，给定了城市的数目和在不同城市间旅行所需要的不同费用。如果要给一个销售员提供一份经过所有城市的路线图，我们的问题是，是否有某种路线是更为廉价的。在这种形

式的问题中，输入数据是n座城市，输出结果则是结论“是”或者“否”。我们需要多少个计算机步骤才能找出廉价的路线呢？一种简单粗暴的计算方法是考虑所有可能的路线。

如果我们面对的是大约100座城市，那么就算有一台来自未来的具有每秒10^{18}次计算能力的超级计算机，用这种方法也还是需要大约4 000个世纪才能解决这一问题。因此，如果能有一种针对该问题的有效算法那就太美妙了，所谓有效算法是只需要多项式时间因而使计算机可以轻松应对的算法。如果有一种像“冒泡法”那样有效的算法，这样的超级计算机能够在不到一微秒的时间里就解决流动销售员问题。

在重视效率的前提下，有两类重要问题是要考虑的，即P和NP问题。

- P问题是指能够在多项式时间内解决的一类问题，例如排序问题。
- NP问题是指能够在多项式时间内得到确认的一类问题。流动销售员问题就是这样的一个问题。尽管我们很难找到流动销售员问题的一个解，但是要通过计算来确认是否有比给定路线更为廉价的路线，却是非常快捷的。两者之间的难度差异恰如在干草堆中找到一根针和确认你已经找到那根针之间的难度差异。

任何一个P问题必然也是NP问题，因为在多项式时间内找到解就是对它的确认。而目前亟待解决的问题是：是否任何一个NP问题也都是P问题？换言之，是否有可能为每个NP问题找到一种多项式时间的算法？如果真是这样，那么就有$P=NP$。另一方面，如果我们能够找到一个NP问题并且证明不可能存在对于它的高效算法，那么P就不可能等于NP。

这就是流动销售员问题具有重要地位的原因。它本身是一个很难的问题，而且所有其他NP类问题都能够转化成它。如果我们能够为这一个问题找到一种高效算法，就可以由此而推出$P=NP$。这个结果意味着所有NP问题都能高效地解决。

如果能够确定$P=NP$，那么由此导出的一个推论是，我们有可能通过一种高效的计算机算法找到一个大数的质因数。现代密码学（参见第14章）所依赖的就是这个问题的难度，它还为电子商务的安全性与计算机网络诚信提供了基础。换言之，一个成功的证明将为黑客与密码窃取者提供一个巨大的竞技场。幸运的是，目前人们认为机器不大可能以高效方式解决所有问题，因而更倾向于认为$P\neq NP$。

数学的未来

为征服这些重大问题而付出的持续努力，无疑将占据数学家们未来很多年时间。解答与证明更有可能是以循序渐进而不是阶段性爆发的方式出现——这是数学发展过程中经常采取的方式，即精雕细刻。概念要经过反复琢磨，不久前才完成的曲折计算今天遭到抛弃，繁复晦涩的论证现在变得多余，通向众多领域的道路变成坦途。与人类思想的任何其他领域一样，时尚与流行趋势也要发挥自己的作用。并不是所有数学内容都得以保留，随着数学理论的修改，大批成果被清除出去。其中一些充其量只能作为历史注记。

不久前才发生的事情能够提示未来发展的线索吗？至少在某种意义上讲，答案确实为“是”。20世纪后半叶迎来了计算机时代，它对数学以及其他很多领域产生了深远的影响。20世纪70年代，由计算机做出的对于“四色”地图猜想的证明（参见第11章）堪称为时代的标志。对于几十年前还无法想象的计算能力，今天的数学家们

可以视为理所当然。时至今日，数学家看起来经常更像是实验室中的科学家。他们可以用计算机来检验各种思想和进行数学实验。计算机不仅能完成算术计算，处理代数问题也不费多少气力，强大的图像处理能力更使我们能够看到很多种几何形状和表面。数学家经常会制作物理模型，但是计算机图像能做得更好。不过我们必须承认，明天的数学家们在拥有越来越丰富的计算资源后，会认为我们今天所具有的能力很弱。

不过毋庸置疑的是，我们今天正处在数学发展的一个好时代中。就在30年前，数学还只能应用于工程和物理学中。今天，她已将双翼伸向了很多新兴领域，而这又导致对于新的数学知识的需求。很明显，计算机科学需要数学，但在过去，人们可能没有料到化学、生物学和心理学（这里只列举3个例子）也会需要借助于数学的投入。实际上，全部数学的疆域是无比广阔的，几乎不会有任何一个学科能够在某个方面完全不受数学影响。

由此导致的后果是，职业数学家是高度专门化的，不仅是专注于数学的某一个分支中，而且要专注于该分支某一领域的一个小角落中。这种精细分化是知识、理论、技术以及数学应用快速增殖的必然后果。几百年前，我们可以想象将世界上最重要的一批数学家聚在一起，他们可以愉快地交谈，（我们假定他们可以使用同一种语言）而且他们能够互相理解，分享同样的知识领域。而在今天，相似的思维劳动可能会使对其自身专长有着深刻领悟的个体分组活动，而大多数人对于其他同事的努力则几乎一无所知。这也许就是进步导致的代价。

未来究竟在何方？本质上讲，数学所需要的是：一个热爱知识的头脑、一支铅笔和一个本子（现在可能是一台电脑）以及想要探索未解之谜的欲望。需要我们去努力的事情还有很多。

术 语 表

π（Pi）
最初出现于和圆有关的问题中，这个最富盛名的常数如今已出现于所有数学领域之中。π的数值近似为$\frac{22}{7}$，或者写成十进小数的话是3.141 592 653 5…

（集合的）基数（Cardinal number）
一个集合中元素的个数。集合$\{a, b, c, d, e\}$的基数是5，不过基数的概念也可以拓展到无穷集合。

$x-y$轴（x–y axes）
这一思想源于笛卡儿在描画点时对其赋予一个x坐标（水平轴）和一个y坐标（垂直轴）。

阿尔冈图（Argand Diagram）
一种能够将二维复数平面呈现出来的可视方法。

八元数，凯莱数（Octonions, Cayley numbers）
19世纪40年代中发现的八维“虚数”。

毕达哥拉斯定理（Pythagoras's theorem）
若一个直角三角形的三边长度为x,y和z，则有$x^2+y^2=z^2$，其中z是与直角相对的最长边（斜边）的长度。

超越数（Transcendental number）
无法表示为诸如$ax^2+bx+c=0$（其中a, b, c为整数；也可能涉及x的更高幂次）等代数方程的解的数。

代数（Algebra）
作为对算术知识的拓展，代数处理的是字母而不是数字，如今它已是一种通用方法，在全部数学知识及其应用领域中发挥作用。“Algebra”这个词来源于公元九世纪时阿拉伯文本中的“*al-jabr*”①。

单位分数（Unit fractions）
上半部分（分子）是1的分数。

定理（Theorem）
这个术语特指具有一定重要性的已确认事实。

对称（Symmetry）
一种形状的规则性。如果这一形状在旋转后可以填入原先的位置中，则称它具有旋转对称性。如果一个图形的镜像能够填入原先的位置，则称它具有镜面对称性。

对偶性（Duality）
两个概念（例如几何学中的点和线以及

① 即解方程时将系数为负的项移动到等号另一边的操作。——译者注

逻辑中的OR和AND）之间的一种相互关系。

多面体（Polyhedron）
一个有很多面的立体图形。例如，一个四面体有4个三角形面，而一个正方体有6个正方形面。

二进制数（Binary number）
以符号0, 1为基来表示的数。二进制数对于计算机运算来说是基础性的。

反比平方律（Inverse square law）
牛顿定律指出，任意两个质量为m_1和m_2的物体，以正比于$\frac{m_1 \times m_2}{d^2}$的力彼此吸引。

斐波那契数列（Fibonacci sequence）
整数数列1, 1, 2, 3, 5, 8, 13, 21, 34, 55, …每一项都是它前面两项的和。

费马最后定理（Fermat's last theorem）
这个定理的内容是，不存在整数值x, y, z使得$x^n + y^n = z^n$，其中n是大于2的整数。

费诺平面（Fano plane）
由7个点和7条线构成的一种几何图形。

分母（Denominator）
一个分数的下半部分。在分数$\frac{3}{7}$中，数字7是分母。

分数（Fraction）
一个整数被另一个除，例如$\frac{3}{7}$。

分形（Fractal）
一种“粗糙的”图形，在任意微观尺度上看起来都是一样的。分形具有自相似的性质。文中图示实例为曼德尔布罗特集。

分子（Numerator）
一个分数的上半部分。在分数$\frac{3}{7}$中，数字3是分子。

复数（Complex number）
形如$a + bi$的数，其中a, b是普通实数，虚数i满足性质$i^2 = -1$。例如复数$2 + 3i$。

概率（Probability）
对于机会的度量，其数值介于0（不可能）到1（必然）之间。

公钥加密（Public key encryption）
将用来给信息加密的密钥公布出来的密码系统。需要解译信息的接受者拥有潜在攻击者无法获得的关于密钥的其他信息。

豪斯道夫维度（Hausdorff dimension）
度量维度的一种方法，它可以给出分数的数值。它对于分形的研究具有特别的适用性。

蝴蝶效应（Butterfly effect）
由彼此非常接近的初始点而产生完全不同的轨迹的效应。

化圆为方（Squaring the circle）
构造与给定圆具有相同面积的正方形的问题——只使用直尺来画直线，用圆规来画圆。这是无法做到的。

黄金比例（Golden ratio）
无处不在的黄金数ϕ（phi），其数值为$\phi = \frac{\sqrt{5}+1}{2} = 1.618\ 033\ 988\ 7\cdots$

混沌理论（Chaos theory）
关于会出现随机现象但又有潜在规律性的动力系统的理论。

积分（Integration）
在微积分中计算面积的一种基本运算。它可以呈现为微分的逆运算。

基（Base）
一个数字系统的基础。古巴比伦人以60为基建立了他们的数系，而我们今天则是以10为基（十进制）。

级数（Series）
一列（可能会有无限多个）数字或符号累加起来。

极限（Limit）
一个含x的表达式，随着x越来越接近于某个特定的数而取得的极限值。

集合（Set）
多个对象的总体。例如，由一些国家的名称组成的集合C={英国，俄罗斯，美国，印度，澳大利亚}，正偶数的集合$E=\{2, 4, 6, 8, \cdots\}$。

几何学（Geometry）
这门学科缘起于公元前3世纪欧几里得的《几何原本》，所讨论的是线、图形和空间的性质。几何学的影响遍及全部数学领域，现在已经丧失了原本有局限性的历史含义。

克利福德代数（Clifford algebra）
一种n维代数，人们发现它对于量子力学非常有用。它的出现归功于数学家W.K.克利福德。

空集（Empty set）
不包含对象的集合。传统上是用$\varnothing$来表示，在集合论中这是一个有用的概念。

零和（Zero-sum）
应用于博弈论。一个玩家所赢得的（+w）就是另一玩家所输掉的（−w），因此其和为零。

零假设（Null hypothesis），H_0
统计学家对于有待检验的假设的称呼。

流形（Manifold）
一种几何对象，本身可能很复杂，但若只考察一小部分的话，就能适用于普通欧氏几何学。例如，炸面圈表面是一个有洞的面，但若只看一小部分，则像是没有洞的普通圆盘一样。

孪生质数（Twin primes）
两个质数，其间隔至多只有一个数：例如11和13是孪生质数。目前尚不清楚是否有无限多对孪生质数。

逻辑斯蒂方程（Logistic equation）
用来对种群从某一年到下一年的增长进行建模的方程：$x_{n+1}=r(1-x_n)x_n$。

欧拉常数e
欧拉常数e是继π之后第二重要的数学常数。尤其是在增长问题中会出现这个数。e的十进制表示数值近似为2.718 281 828 5…

欧氏几何学（Euclid's geometry）
传统几何学，呈现于公元前300年时欧几里得的《几何原本》之中。在这种几何学中，一个三角形的内角和等于180°。

平方根（Square root）
自乘后能够得到给定数的数。例如3是9的平方根，因为3×3＝9。

平方数（Square number）
将一个整数自乘所得到的结果。数9是一个平方数，因为9＝3×3。平方数包括：1, 4, 9, 16, 25, 36, 49, 64, …

球面（Sphere）
一个正常球面是球的二维表面。它的方程中包含x, y, z三个变量，$x^2+y^2+z^2=1$。一个n维球面具有类似的包含$(n+1)$个变量的方程。

群（Group）
一种由若干元素组成的数学结构，这些元素能够结合而产生出同种类型的元素，并且具有特定的性质：所有元素都有一个逆元素，并且运算遵循结合律$a\times(b\times c)=(a\times b)\times c$。

三体问题（Three-body problem）
对三个互相吸引的物体（例如月球、地球和太阳）的运动过程进行分析的努力。

十进制系统（Decimal system）
我们通常使用的以0, 1, …, 9这十个符号来表示基的数系。

十六进制数（Hexadecimal number）
一种用符号0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E和F来表示16个基的数系。它在计算中有着广泛应用。

数列（Sequence）
一列（可能会有无限多个）数字或符号。

双曲几何（Hyperbolic geometry）
一种非欧几何学，其中三角形的内角和小于180°。

四元数（Quaternions）
19世纪40年代中由W. R. 哈密顿所发现的四维“虚数”。

算法（Algorithm）
一种数学技巧；用来解决问题的一套规则。

椭圆几何（Elliptical geometry）
一种非欧几何，其中三角形的内角和大于180°。

拓扑学（Topology）
也称为“橡胶板”几何学，主要关注点包括曲面在因拉伸或压缩（但不包括剪切）而变形时保持不变的性质等。在拓扑学中，长度和角度的度量失去了它们在普通几何学中所具有的地位。

微分（Differentiation）
在微积分中用来产生导数或变化率的基本运算。例如，对于一个描述距离随着时间变化的表达式来说，导数代表速度。速度表达式的导数代表加速度。

微分方程（Differential equation）
涉及速度、加速度和其他变化率等微积分概念的方程。重要的实例包括麦克斯韦方程（组）和纳维尔-斯托克斯方程（组）。

位值制系统（Place-value system）一个数的大小取决于其中数字的位置。在73中，7的位值表示“7个十”，而3的位值表示“3个单位一”。

无理数（Irrational numbers）
无法表示为分数形式的数（例如2的平方根）。

吸引子（Attractor）
在动力系统中吸引运动曲线的点。对于逐渐停止的摆来说，运动曲线或轨道被吸引向一个单独的点，在这个点处位移是0，速度也是0。对于复杂的动力系统，吸引子可能会构成一个分形，在这种情况下称其为“奇异吸引子”。

镶嵌（Tessellation）
用一种或多种给定形状填充平坦空间使得各个形状之间刚好彼此吻合的方法。例如，只用正六边形可以形成一种镶嵌，而只用正五边形则不行。

虚数（Imaginary numbers）
涉及“虚数”$i=\sqrt{-1}$的数。在与普通的数（或称“实数”）结合起来后，能够形成复数。

一一对应（One-to-one correspondence）
一种关系类型，指一个集合中的每一个对象恰好对应于另一个集合中的一个对象，反之亦然。

伊尼格玛（Enigma）
一种复杂的电子机械装置，在第二次世界大战中用来给信息加密。

有理数（Rational numbers）
要么是整数要么是分数的数。

余数（Remainder）
如果将一个整数用另一个整数去除，那么剩下的数就是余数。例如17被3除，商是5，余数是2。

指数（Exponent）
算术中使用的一种符号。将一个数自乘，例如5×5记为指数为2的形式，即5^2。表达式$5\times5\times5$记为5^3，依次类推。这个符号还可以拓展：例如数$5^{1/2}$意味着5的平方根。与它含义相同的词是乘方和幂。

质数（Prime number）
不能被任何数（除了自身和1之外）除尽的整数。例如，7是一个质数，而6则不是（因为6可以被2除尽）；2是最小的质数。

版权声明